Preface

Welcome to *College Physics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it’s our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 20 books for college and AP courses used by hundreds of thousands of students. Our adaptive learning technology, designed to improve learning outcomes through personalized educational paths, is being piloted in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax Resources

Customization

College Physics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit your book page on openstax.org for more information.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on openstax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on openstax.org.

Format

You can access this textbook for free in web view or PDF through openstax.org, and in low-cost print and iBooks editions.

About *College Physics*

College Physics meets standard scope and sequence requirements for a two-semester introductory algebra-based physics course. The text is grounded in real-world examples to help students grasp fundamental physics concepts. It requires knowledge of algebra and some trigonometry, but not calculus. *College Physics* includes learning objectives, concept questions, links to labs and simulations, and ample practice opportunities for traditional physics application problems.

Coverage and Scope

College Physics is organized such that topics are introduced conceptually with a steady progression to precise definitions and analytical applications. The analytical aspect (problem solving) is tied back to the conceptual before moving on to another topic. Each introductory chapter, for example, opens with an engaging photograph relevant to the subject of the chapter and interesting applications that are easy for most students to visualize.

Concepts and Calculations

The ability to calculate does not guarantee conceptual understanding. In order to unify conceptual, analytical, and calculation skills within the learning process, we have integrated Strategies and Discussions throughout the text.

Modern Perspective

The chapters on modern physics are more complete than many other texts on the market, with an entire chapter devoted to medical applications of nuclear physics and another to particle physics. The final chapter of the text, “Frontiers of Physics,” is devoted to the most exciting endeavors in physics. It ends with a module titled “Some Questions We Know to Ask.”

Key Features

Modularity

This textbook is organized as a collection of modules that can be rearranged and modified to suit the needs of a particular professor or class. That being said, modules often contain references to content in other modules, as most topics in physics cannot be discussed in isolation.

Learning Objectives

Every module begins with a set of learning objectives. These objectives are designed to guide the instructor in deciding what content to include or assign, and to guide the student with respect to what he or she can expect to learn. After completing the module and end-of-module exercises, students should be able to demonstrate mastery of the learning objectives.

Call-Outs

Key definitions, concepts, and equations are called out with a special design treatment. Call-outs are designed to catch readers’ attention, to make it clear that a specific term, concept, or equation is particularly important, and to provide easy reference for a student reviewing content.

Key Terms

Key terms are in bold and are followed by a definition in context. Definitions of key terms are also listed in the Glossary, which appears at the end of the module.

Worked Examples

Worked examples have four distinct parts to promote both analytical and conceptual skills. Worked examples are introduced in words, always using some application that should be of interest. This is followed by a Strategy section that emphasizes the concepts involved and how solving the problem relates to those concepts. This is followed by the mathematical Solution and Discussion.

Many worked examples contain multiple-part problems to help the students learn how to approach normal situations, in which problems tend to have multiple parts. Finally, worked examples employ the techniques of the problem-solving strategies so that students can see how those strategies succeed in practice as well as in theory.

Problem-Solving Strategies

Problem-solving strategies are first presented in a special section and subsequently appear at crucial points in the text where students can benefit most from them. Problem-solving strategies have a logical structure that is reinforced in the worked examples and supported in certain places by line drawings that illustrate various steps.

Misconception Alerts

Students come to physics with preconceptions from everyday experiences and from previous courses. Some of these preconceptions are misconceptions, and many are very common among students and the general public. Some are inadvertently picked up through misunderstandings of lectures and texts. The Misconception Alerts feature is designed to point these out and correct them explicitly.

Take-Home Investigations

Take Home Investigations provide the opportunity for students to apply or explore what they have learned with a hands-on activity.

Things Great and Small

In these special topic essays, macroscopic phenomena (such as air pressure) are explained with submicroscopic phenomena (such as atoms bouncing off walls). These essays support the modern perspective by describing aspects of modern physics before they are formally treated in later chapters. Connections are also made between apparently disparate phenomena.

Simulations

Where applicable, students are directed to the interactive PHeT physics simulations developed by the University of Colorado. There they can further explore the physics concepts they have learned about in the module.

Summary

Module summaries are thorough and functional and present all important definitions and equations. Students are able to find the definitions of all terms and symbols as well as their physical relationships. The structure of the summary makes plain the fundamental principles of the module or collection and serves as a useful study guide.

Glossary

At the end of every module or chapter is a Glossary containing definitions of all of the key terms in the module or chapter.

End-of-Module Problems

At the end of every chapter is a set of Conceptual Questions and/or skills-based Problems & Exercises. Conceptual Questions challenge students' ability to explain what they have learned conceptually, independent of the mathematical details. Problems & Exercises challenge students to apply both concepts and skills to solve mathematical physics problems. Online, every other problem includes an answer that students can reveal immediately by clicking on a “Show Solution” button.

In addition to traditional skills-based problems, there are three special types of end-of-module problems: Integrated Concept Problems, Unreasonable Results Problems, and Construct Your Own Problems. All of these problems are indicated with a subtitle preceding the problem.

Integrated Concept Problems

In Integrated Concept Problems, students are asked to apply what they have learned about two or more concepts to arrive at a solution to a problem. These problems require a higher level of thinking because, before solving a problem, students have to recognize the combination of strategies required to solve it.

Unreasonable Results

In Unreasonable Results Problems, students are challenged to not only apply concepts and skills to solve a problem, but also to analyze the answer with respect to how likely or realistic it really is. These problems contain a premise that produces an unreasonable answer and are designed to further emphasize that properly applied physics must describe nature accurately and is not simply the process of solving equations.

Construct Your Own Problem

These problems require students to construct the details of a problem, justify their starting assumptions, show specific steps in the problem’s solution, and finally discuss the meaning of the result. These types of problems relate well to both conceptual and analytical aspects of physics, emphasizing that physics must describe nature. Often they involve an integration of topics from more than one chapter. Unlike other problems, solutions are not provided since there is no single correct answer.

Instructors should feel free to direct students regarding the level and scope of their considerations. Whether the problem is solved and described correctly will depend on initial assumptions.

Additional Resources

Student and Instructor Resources

We’ve compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which can be requested on your openstax.org log-in. Take advantage of these resources to supplement your OpenStax book.

Partner Resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on openstax.org.

About the Authors

Senior Contributing Authors

Paul Peter Urone, Professor Emeritus at California State University, Sacramento
Roger Hinrichs, State University of New York, College at Oswego

Contributing Authors

Kim Dirks, University of Auckland
Manjula Sharma, University of Sydney

Reviewers

Matthew Adams, Crafton Hills College, San Bernardino Community College District
Erik Christensen, South Florida Community College
Douglas Ingram, Texas Christian University
Eric Kincanon, Gonzaga University
Lee H. LaRue, Paris Junior College
Chuck Pearson, Virginia Intermont College
Marc Sher, College of William and Mary
Ulrich Zurcher, Cleveland State University

Introduction to Science and the Realm of Physics, Physical Quantities, and Units
class="introduction"

Galaxies are as immense as atoms are small. Yet the same laws of physics describe both, and all the rest of nature—an indication of the underlying unity in the universe. The laws of physics are surprisingly few in number, implying an underlying simplicity to nature’s apparent complexity. (credit: NASA, JPL-Caltech, P. Barmby, Harvard-Smithsonian Center for

Astrophysics
)

What is your first reaction when you hear the word “physics”? Did you imagine working through difficult equations or memorizing formulas that seem to have no real use in life outside the physics classroom? Many people come to the subject of physics with a bit of fear. But as you begin your exploration of this broad-ranging subject, you may soon come to realize that physics plays a much larger role in your life than you first thought, no matter your life goals or career choice.

For example, take a look at the image above. This image is of the Andromeda Galaxy, which contains billions of individual stars, huge clouds of gas, and dust. Two smaller galaxies are also visible as bright blue spots in the background. At a staggering 2.5 million light years from the Earth, this galaxy is the nearest one to our own galaxy (which is called the Milky Way). The stars and planets that make up Andromeda might seem to be the furthest thing from most people’s regular, everyday lives. But Andromeda is a great starting point to think about the forces that hold together the universe. The forces that cause Andromeda to act as it does are the same forces we contend with here on Earth, whether we are planning to send a rocket into space or simply raise the walls for a new home. The same gravity that causes the stars of Andromeda to rotate and revolve also causes water to flow over hydroelectric dams here on Earth. Tonight, take a moment to look up at the stars. The forces out there are the same as the ones here on Earth. Through a study of physics, you may gain a greater understanding of the interconnectedness of everything we can see and know in this universe.

Think now about all of the technological devices that you use on a regular basis. Computers, smart phones, GPS systems, MP3 players, and satellite radio might come to mind. Next, think about the most exciting modern technologies that you have heard about in the news, such as trains that levitate above tracks, “invisibility cloaks” that bend light around them, and microscopic robots that fight cancer cells in our bodies. All of these groundbreaking advancements, commonplace or unbelievable, rely on the principles of physics. Aside from playing a significant role in technology, professionals such as engineers, pilots, physicians, physical therapists, electricians, and computer programmers apply physics concepts in their daily work. For example, a pilot must understand how wind forces affect a flight path and a physical therapist must understand how the muscles in the body experience forces as they move and bend. As you will learn in this text, physics principles are propelling new, exciting technologies, and these principles are applied in a wide range of careers.

In this text, you will begin to explore the history of the formal study of physics, beginning with natural philosophy and the ancient Greeks, and leading up through a review of Sir Isaac Newton and the laws of physics that bear his name. You will also be introduced to the standards scientists use when they study physical quantities and the interrelated system of measurements most of the scientific community uses to communicate in a single mathematical language. Finally, you will study the limits of our ability to be accurate and precise, and the reasons scientists go to painstaking lengths to be as clear as possible regarding their own limitations.

Physics: An Introduction

- Explain the difference between a principle and a law.
- Explain the difference between a model and a theory.

The flight formations of migratory birds such as Canada geese are governed by the laws of physics. (credit: David Merrett)

The physical universe is enormously complex in its detail. Every day, each of us observes a great variety of objects and phenomena. Over the centuries, the curiosity of the human race has led us collectively to explore and catalog a tremendous wealth of information. From the flight of birds to the colors of flowers, from lightning to gravity, from quarks to clusters of galaxies, from the flow of time to the mystery of the creation of the universe, we have asked questions and assembled huge arrays of facts. In the face of all these details, we have discovered that a surprisingly small and unified set of physical laws can explain what we observe. As humans, we make generalizations and seek order. We have found that nature is remarkably cooperative—it exhibits the *underlying order and simplicity* we so value.

It is the underlying order of nature that makes science in general, and physics in particular, so enjoyable to study. For example, what do a bag of chips and a car battery have in common? Both contain energy that can be converted to other forms. The law of conservation of energy (which says that energy can change form but is never lost) ties together such topics as food calories, batteries, heat, light, and watch springs. Understanding this law makes it easier to learn about the various forms energy takes and how they relate to one another. Apparently unrelated topics are connected through broadly applicable physical laws, permitting an understanding beyond just the memorization of lists of facts.

The unifying aspect of physical laws and the basic simplicity of nature form the underlying themes of this text. In learning to apply these laws, you will, of course, study the most important topics in physics. More importantly, you will gain analytical abilities that will enable you to apply these laws far beyond the scope of what can be included in a single book. These analytical skills will help you to excel academically, and they will also help you to think critically in any professional career you choose to pursue. This module discusses the realm of physics (to define what physics is), some applications of physics (to illustrate its relevance to other disciplines), and more precisely what constitutes a physical law (to illuminate the importance of experimentation to theory).

Science and the Realm of Physics

Science consists of the theories and laws that are the general truths of nature as well as the body of knowledge they encompass. Scientists are continually trying to expand this body of knowledge and to perfect the expression of the laws that describe it. **Physics** is concerned with describing the interactions of energy, matter, space, and time, and it is especially interested in what fundamental mechanisms underlie every phenomenon. The concern for describing the basic phenomena in nature essentially defines the *realm of physics*.

Physics aims to describe the function of everything around us, from the movement of tiny charged particles to the motion of people, cars, and spaceships. In fact, almost everything around you can be described quite accurately by the laws of physics. Consider a smart phone ([link]). Physics describes how electricity interacts with the various circuits inside the device. This knowledge helps engineers select the appropriate materials and circuit layout when building the smart phone. Next, consider a GPS system. Physics describes the relationship between the speed of an object, the distance over which it travels, and the time it takes to travel that distance. When you use a GPS device in a vehicle, it utilizes these physics equations to determine the travel time from one location to another.

The Apple "iPhone" is a common smart phone with a GPS function. Physics describes the way that electricity flows through the circuits of this device. Engineers use their knowledge of physics to construct an iPhone with features that consumers will enjoy. One specific feature of an iPhone is the GPS function. GPS uses physics equations to determine the driving time between two locations on a map. (credit: @gletham GIS, Social, Mobile Tech Images)

Applications of Physics

You need not be a scientist to use physics. On the contrary, knowledge of physics is useful in everyday situations as well as in nonscientific professions. It can help you understand how microwave ovens work, why metals should not be put into them, and why they might affect pacemakers. (See [link] and [link].) Physics allows you to understand the hazards of radiation and rationally evaluate these hazards more easily. Physics also explains the reason why a black car radiator helps remove heat in a car engine, and it explains why a white roof helps keep the inside of a house cool. Similarly, the operation of a car's ignition system as well as the transmission of electrical signals through our body's nervous system are much easier to understand when you think about them in terms of basic physics.

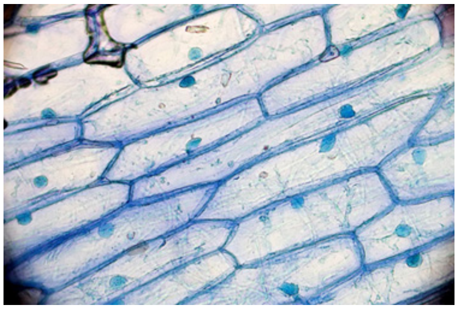
Physics is the foundation of many important disciplines and contributes directly to others. Chemistry, for example—since it deals with the interactions of atoms and molecules—is rooted in atomic and molecular physics. Most branches of engineering are applied physics. In architecture, physics is at the heart of structural stability, and is involved in the acoustics, heating, lighting, and cooling of buildings. Parts of geology rely heavily on physics, such as radioactive dating of rocks, earthquake analysis, and heat transfer in the Earth. Some disciplines, such as biophysics and geophysics, are hybrids of physics and other disciplines.

Physics has many applications in the biological sciences. On the microscopic level, it helps describe the properties of cell walls and cell membranes ([link] and [link]). On the macroscopic level, it can explain the heat, work, and power associated with the human body. Physics is involved in medical diagnostics, such as x-rays, magnetic resonance imaging (MRI), and ultrasonic blood flow measurements. Medical therapy sometimes directly involves physics; for example, cancer radiotherapy uses ionizing radiation. Physics can also explain sensory phenomena, such as how musical instruments make sound, how the eye detects color, and how lasers can transmit information.

It is not necessary to formally study all applications of physics. What is most useful is knowledge of the basic laws of physics and a skill in the analytical methods for applying them. The study of physics also can improve your problem-solving skills. Furthermore, physics has retained the most basic aspects of science, so it is used by all of the sciences, and the study of physics makes other sciences easier to understand.

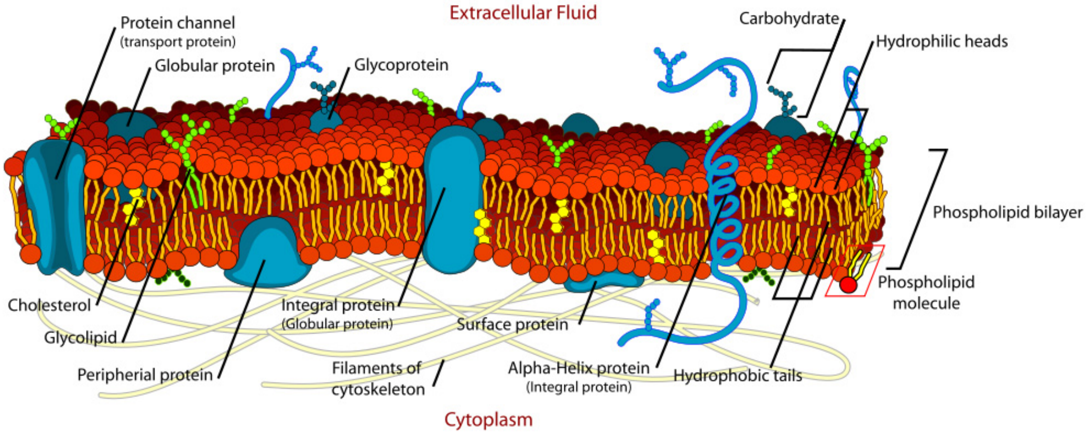
The laws of physics help us understand how common appliances work. For example, the laws of physics can help explain how microwave ovens heat up food, and they also help us understand why it is dangerous to place metal objects in a microwave oven. (credit: MoneyBlogNewz)

These two applications of physics have more in common than meets the eye. Microwave ovens use electromagnetic waves to heat food. Magnetic resonance imaging (MRI) also uses electromagnetic waves to yield an image of the brain, from which the exact location of tumors can be determined. (credit: Rashmi Chawla, Daniel Smith, and Paul E. Marik)



Physics, chemistry,

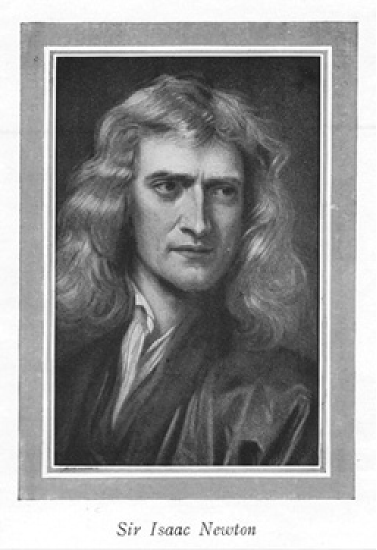
and biology help describe the properties of cell walls in plant cells, such as the onion cells seen here. (credit: Umberto Salvagnin)



An artist’s rendition of the the structure of a cell membrane. Membranes form the boundaries of animal cells and are complex in structure and function. Many of the most fundamental properties of life, such as the firing of nerve cells, are related to membranes. The disciplines of biology, chemistry, and physics all help us understand the membranes of animal cells. (credit: Mariana Ruiz)

Models, Theories, and Laws; The Role of Experimentation

The laws of nature are concise descriptions of the universe around us; they are human statements of the underlying laws or rules that all natural processes follow. Such laws are intrinsic to the universe; humans did not create them and so cannot change them. We can only discover and understand them. Their discovery is a very human endeavor, with all the elements of mystery, imagination, struggle, triumph, and disappointment inherent in any creative effort. (See [link] and [link].) The cornerstone of discovering natural laws is observation; science must describe the universe as it is, not as we may imagine it to be.



Isaac Newton (1642–1727) was very reluctant to publish his revolutionary work and had to be convinced to do so. In his later years, he stepped down from his academic post and became exchequer of the Royal Mint. He took this post seriously, inventing reeding (or creating ridges) on the edge of coins to prevent unscrupulous people from trimming the silver off of them before using them as currency. (credit: Arthur Shuster and Arthur E. Shipley: *Britain's Heritage of Science*. London, 1917.)

Marie Curie (1867–1934) sacrificed

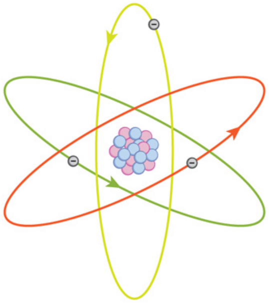
monetary assets to help finance her early research and damaged her physical well-being with radiation exposure. She is the only person to win Nobel prizes in both physics and chemistry. One of her daughters also won a Nobel Prize. (credit: Wikimedia Commons)

We all are curious to some extent. We look around, make generalizations, and try to understand what we see—for example, we look up and wonder whether one type of cloud signals an oncoming storm. As we become serious about exploring nature, we become more organized and formal in collecting and analyzing data. We attempt greater precision, perform controlled experiments (if we can), and write down ideas about how the data may be organized and unified. We then formulate models, theories, and laws based on the data we have collected and analyzed to generalize and communicate the results of these experiments.

A **model** is a representation of something that is often too difficult (or impossible) to display directly. While a model is justified with experimental proof, it is only accurate under limited situations. An example is the planetary model of the atom in which electrons are pictured as orbiting the nucleus, analogous to the way planets orbit the Sun. (See [link].) We cannot observe electron orbits directly, but the mental image helps explain the observations we can make, such as the emission of light from hot gases (atomic spectra). Physicists use models for a variety of purposes. For example, models can help physicists analyze a scenario and perform a calculation, or they can be used to represent a situation in the form of a computer simulation. A **theory** is an explanation for patterns in nature that is supported by scientific evidence and verified multiple times by various groups of researchers. Some theories include models to help visualize phenomena, whereas others do not. Newton’s theory of gravity, for example, does not require a model or mental image, because we can observe the objects directly with our own senses. The kinetic theory of gases, on the other hand, is a model in which a gas is viewed as being composed of atoms and molecules. Atoms and molecules are too small to be observed directly with our senses—thus, we picture them mentally to understand what our instruments tell us about the behavior of gases.

A **law** uses concise language to describe a generalized pattern in nature that is supported by scientific evidence and repeated experiments. Often, a law can be expressed in the form of a single mathematical equation. Laws and theories are similar in that they are both scientific statements that result from a tested hypothesis and are supported by scientific evidence. However, the designation *law* is reserved for a concise and very general statement that describes phenomena in nature, such as the law that energy is conserved during any process, or Newton’s second law of motion, which relates force, mass, and acceleration by the simple equation $\mathbf{F} = m\mathbf{a}$. A theory, in contrast, is a less concise statement of observed phenomena. For example, the Theory of Evolution and the Theory of Relativity cannot be expressed concisely enough to be considered a law. The biggest difference between a law and a theory is that a theory is much more complex and dynamic. A law describes a single action, whereas a theory explains an entire group of related phenomena. And, whereas a law is a postulate that forms the foundation of the scientific method, a theory is the end result of that process.

Less broadly applicable statements are usually called principles (such as Pascal’s principle, which is applicable only in fluids), but the distinction between laws and principles often is not carefully made.



What is a model? This planetary model of the atom shows electrons orbiting the nucleus. It is a drawing that we use to form a mental image of the atom that we cannot see directly with our eyes because it is too small.

Note:

Models, Theories, and Laws

Models, theories, and laws are used to help scientists analyze the data they have already collected. However, often after a model, theory, or law has been developed, it points scientists toward new discoveries they would not otherwise have made.

The models, theories, and laws we devise sometimes *imply the existence of objects or phenomena as yet unobserved.* These predictions are remarkable triumphs and tributes to the power of science. It is the underlying order in the universe that enables scientists to make such spectacular predictions. However, if *experiment* does not verify our predictions, then the theory or law is wrong, no matter how elegant or convenient it is. Laws can never be known with absolute certainty because it is impossible to perform every imaginable experiment in order to confirm a law in every possible scenario. Physicists operate under the assumption that all scientific laws and theories are valid until a counterexample is observed. If a good-quality, verifiable experiment contradicts a well-established law, then the law must be modified or overthrown completely.

The study of science in general and physics in particular is an adventure much like the exploration of uncharted ocean. Discoveries are made; models, theories, and laws are formulated; and the beauty of the physical universe is made more sublime for the insights gained.

Note:

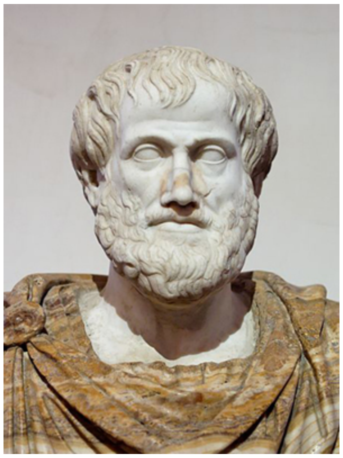
The Scientific Method

As scientists inquire and gather information about the world, they follow a process called the **scientific method**. This process typically begins with an observation and question that the scientist will research. Next, the scientist typically performs some research about the topic and then devises a hypothesis. Then, the scientist will test the hypothesis by performing an experiment. Finally, the scientist analyzes the results of the experiment and draws a conclusion. Note that the scientific method can be applied to many situations that are not limited to science, and this method can be modified to suit the situation.

Consider an example. Let us say that you try to turn on your car, but it will not start. You undoubtedly wonder: Why will the car not start? You can follow a scientific method to answer this question. First off, you may perform some research to determine a variety of reasons why the car will not start. Next, you will state a hypothesis. For example, you may believe that the car is not starting because it has no engine oil. To test this, you open the hood of the car and examine the oil level. You observe that the oil is at an acceptable level, and you thus conclude that the oil level is not contributing to your car issue. To troubleshoot the issue further, you may devise a new hypothesis to test and then repeat the process again.

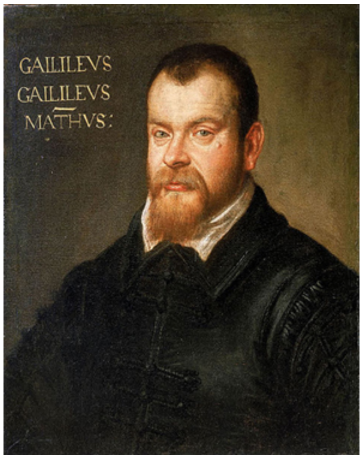
The Evolution of Natural Philosophy into Modern Physics

Physics was not always a separate and distinct discipline. It remains connected to other sciences to this day. The word *physics* comes from Greek, meaning nature. The study of nature came to be called “natural philosophy.” From ancient times through the Renaissance, natural philosophy encompassed many fields, including astronomy, biology, chemistry, physics, mathematics, and medicine. Over the last few centuries, the growth of knowledge has resulted in ever-increasing specialization and branching of natural philosophy into separate fields, with physics retaining the most basic facets. (See [link], [link], and [link].) Physics as it developed from the Renaissance to the end of the 19th century is called **classical physics**. It was transformed into modern physics by revolutionary discoveries made starting at the beginning of the 20th century.

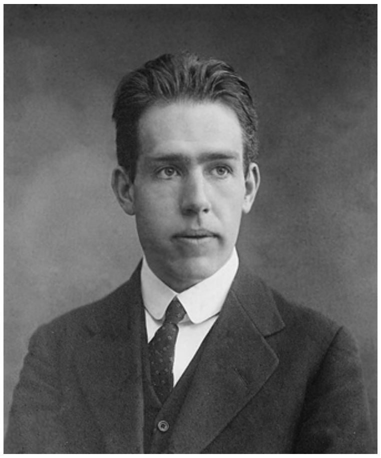


Over the centuries, natural philosophy has evolved into more specialized disciplines, as illustrated by the contributions of some of the greatest minds in history. The Greek philosopher **Aristotle** (384–322 B.C.) wrote on a broad range of topics including physics, animals, the soul, politics, and poetry. (credit: Jastrow

(2006)/Ludovisi Collection)



Galileo Galilei (1564–1642) laid the foundation of modern experimentation and made contributions in mathematics, physics, and astronomy. (credit: Domenico Tintoretto)



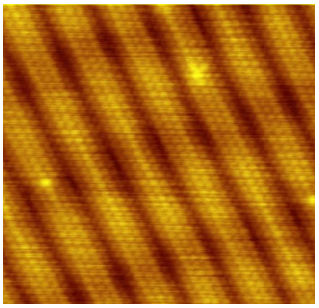
Niels Bohr (1885–1962) made fundamental contributions to the development of quantum mechanics, one part of modern physics. (credit: United States Library of Congress Prints and Photographs Division)

Classical physics is not an exact description of the universe, but it is an excellent approximation under the following conditions: Matter must be moving at speeds less than about 1% of the speed of light, the objects dealt with must be large enough to be seen with a microscope, and only weak gravitational fields, such as the field generated by the Earth, can be involved. Because humans live under such circumstances, classical physics seems intuitively reasonable, while many aspects of modern physics seem bizarre. This is why models are so useful in modern physics—they let us conceptualize phenomena we do not ordinarily experience. We can relate to models in human terms and visualize what happens when objects move at high speeds or imagine what objects too small to observe with our senses might be like. For example, we can understand an atom's properties because we can picture it in our minds, although we have never seen an atom with our eyes. New tools, of course, allow us to better picture phenomena we cannot see. In fact, new instrumentation has allowed us in recent years to actually "picture" the atom.

Note:

Limits on the Laws of Classical Physics

For the laws of classical physics to apply, the following criteria must be met: Matter must be moving at speeds less than about 1% of the speed of light, the objects dealt with must be large enough to be seen with a microscope, and only weak gravitational fields (such as the field generated by the Earth) can be involved.



Using a scanning tunneling microscope (STM), scientists can see the individual atoms that compose this sheet of gold. (credit: Erwinrossen)

Some of the most spectacular advances in science have been made in modern physics. Many of the laws of classical physics have been modified or rejected, and revolutionary changes in technology, society, and our view of the universe have resulted. Like science fiction, modern physics is filled with fascinating objects beyond our normal experiences, but it has the advantage over science fiction of being very real. Why, then, is the majority of this text devoted to topics of classical physics? There are two main reasons: Classical physics gives an extremely accurate description of the universe under a wide range of everyday circumstances, and knowledge of classical physics is necessary to understand modern physics.

Modern physics itself consists of the two revolutionary theories, relativity and quantum mechanics. These theories deal with the very fast and the very small, respectively. **Relativity** must be used whenever an object is traveling at greater than about 1% of the speed of light or experiences a strong gravitational field such as that near the Sun. **Quantum mechanics** must be used for objects smaller than can be seen with a microscope. The combination of these two theories is *relativistic quantum mechanics*, and it describes the behavior of small objects traveling at high speeds or experiencing a strong gravitational field. Relativistic quantum mechanics is the best universally applicable theory we have. Because of its mathematical complexity, it is used only when necessary, and the other theories are used whenever they will produce sufficiently accurate results. We will find, however, that we can do a great deal of modern physics with the algebra and trigonometry used in this text.

Exercise:

Check Your Understanding

Problem:

A friend tells you he has learned about a new law of nature. What can you know about the information even before your friend describes the law? How would the information be different if your friend told you he had learned about a scientific theory rather than a law?

Solution:

Without knowing the details of the law, you can still infer that the information your friend has learned conforms to the requirements of all laws of nature: it will be a concise description of the universe around us; a statement of the underlying rules that all natural processes follow. If the information had been a theory, you would be able to infer that the information will be a large-scale, broadly applicable generalization.

Note:
PhET Explorations: Equation Grapher
Learn about graphing polynomials. The shape of the curve changes as the constants are adjusted. View the curves for the individual terms (e.g. $y = bx$) to see how they add to generate the polynomial curve.
https://phet.colorado.edu/sims/equation-grapher/equation-grapher_en.html

Summary

- Science seeks to discover and describe the underlying order and simplicity in nature.
- Physics is the most basic of the sciences, concerning itself with energy, matter, space and time, and their interactions.
- Scientific laws and theories express the general truths of nature and the body of knowledge they encompass. These laws of nature are rules that all natural processes appear to follow.

Conceptual Questions

Exercise:

Problem:

Models are particularly useful in relativity and quantum mechanics, where conditions are outside those normally encountered by humans. What is a model?

Exercise:

Problem: How does a model differ from a theory?

Exercise:

Problem:

If two different theories describe experimental observations equally well, can one be said to be more valid than the other (assuming both use accepted rules of logic)?

Exercise:

Problem: What determines the validity of a theory?

Exercise:

Problem:

Certain criteria must be satisfied if a measurement or observation is to be believed. Will the criteria necessarily be as strict for an expected result as for an unexpected result?

Exercise:

Problem:

Can the validity of a model be limited, or must it be universally valid? How does this compare to the required validity of a theory or a law?

Exercise:

Problem:

Classical physics is a good approximation to modern physics under certain circumstances. What are they?

Exercise:

Problem: When is it *necessary* to use relativistic quantum mechanics?

Exercise:

Problem:

Can classical physics be used to accurately describe a satellite moving at a speed of 7500 m/s? Explain why or why not.

Glossary

classical physics
: physics that was developed from the Renaissance to the end of the 19th century

physics
: the science concerned with describing the interactions of energy, matter, space, and time; it is especially interested in what fundamental mechanisms underlie every phenomenon

model
: representation of something that is often too difficult (or impossible) to display directly

theory
: an explanation for patterns in nature that is supported by scientific evidence and verified multiple times by various groups of researchers

law
: a description, using concise language or a mathematical formula, a generalized pattern in nature that is supported by scientific evidence

and repeated experiments

scientific method

a method that typically begins with an observation and question that the scientist will research; next, the scientist typically performs some research about the topic and then devises a hypothesis; then, the scientist will test the hypothesis by performing an experiment; finally, the scientist analyzes the results of the experiment and draws a conclusion

modern physics

the study of relativity, quantum mechanics, or both

relativity

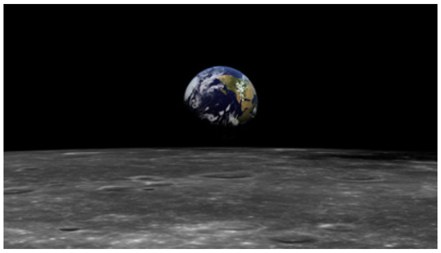
the study of objects moving at speeds greater than about 1% of the speed of light, or of objects being affected by a strong gravitational field

quantum mechanics

the study of objects smaller than can be seen with a microscope

Physical Quantities and Units

- Perform unit conversions both in the SI and English units.
- Explain the most common prefixes in the SI units and be able to write them in scientific notation.

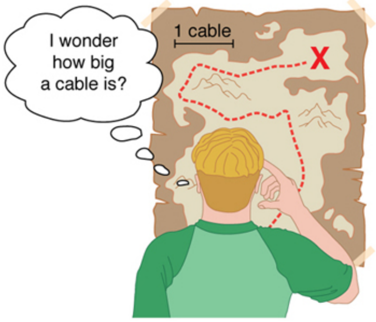


The distance from Earth to the Moon may seem immense, but it is just a tiny fraction of the distances from Earth to other celestial bodies. (credit: NASA)

The range of objects and phenomena studied in physics is immense. From the incredibly short lifetime of a nucleus to the age of the Earth, from the tiny sizes of sub-nuclear particles to the vast distance to the edges of the known universe, from the force exerted by a jumping flea to the force between Earth and the Sun, there are enough factors of 10 to challenge the imagination of even the most experienced scientist. Giving numerical values for physical quantities and equations for physical principles allows us to understand nature much more deeply than does qualitative description alone. To comprehend these vast ranges, we must also have accepted units in which to express them. And we shall find that (even in the potentially mundane discussion of meters, kilograms, and seconds) a profound simplicity of nature appears—all physical quantities can be expressed as combinations of only four fundamental physical quantities: length, mass, time, and electric current.

We define a **physical quantity** either by *specifying how it is measured* or by *stating how it is calculated* from other measurements. For example, we define distance and time by specifying methods for measuring them, whereas we define *average speed* by stating that it is calculated as distance traveled divided by time of travel.

Measurements of physical quantities are expressed in terms of **units**, which are standardized values. For example, the length of a race, which is a physical quantity, can be expressed in units of meters (for sprinters) or kilometers (for distance runners). Without standardized units, it would be extremely difficult for scientists to express and compare measured values in a meaningful way. (See [link].)



Distances given in unknown units are maddeningly useless.

There are two major systems of units used in the world: **SI units** (also known as the metric system) and **English units** (also known as the customary or imperial system). **English units** were historically used in nations once ruled by the British Empire and are still widely used in the United States. Virtually every other country in the world now uses SI units as the standard; the metric system is also the standard system agreed upon by scientists and mathematicians. The acronym “SI” is derived from the French *Système International*.

SI Units: Fundamental and Derived Units

[link] gives the fundamental SI units that are used throughout this textbook. This text uses non-SI units in a few applications where they are in very common use, such as the measurement of blood pressure in millimeters of mercury (mm Hg). Whenever non-SI units are discussed, they will be tied to SI units through conversions.

Length	Mass	Time	Electric Current
meter (m)	kilogram (kg)	second (s)	ampere (A)

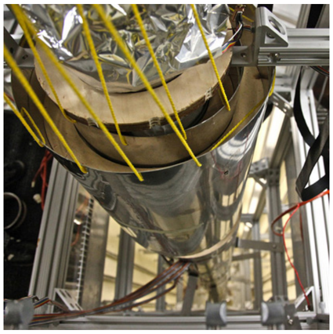
Fundamental SI Units

It is an intriguing fact that some physical quantities are more fundamental than others and that the most fundamental physical quantities can be defined *only* in terms of the procedure used to measure them. The units in which they are measured are thus called **fundamental units**. In this textbook, the fundamental physical quantities are taken to be length, mass, time, and electric current. (Note that electric current will not be introduced until much later in this text.) All other physical quantities, such as force and electric charge, can be expressed as algebraic combinations of length, mass, time, and current (for example, speed is length divided by time); these units are called **derived units**.

Units of Time, Length, and Mass: The Second, Meter, and Kilogram

The Second

The SI unit for time, the **second**(abbreviated s), has a long history. For many years it was defined as 1/86,400 of a mean solar day. More recently, a new standard was adopted to gain greater accuracy and to define the second in terms of a non-varying, or constant, physical phenomenon (because the solar day is getting longer due to very gradual slowing of the Earth’s rotation). Cesium atoms can be made to vibrate in a very steady way, and these vibrations can be readily observed and counted. In 1967 the second was redefined as the time required for 9,192,631,770 of these vibrations. (See [link].) Accuracy in the fundamental units is essential, because all measurements are ultimately expressed in terms of fundamental units and can be no more accurate than are the fundamental units themselves.



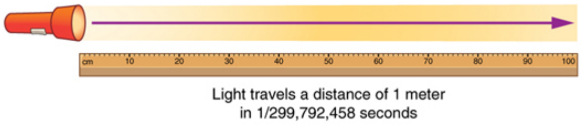
An atomic clock such as this one uses the vibrations of cesium atoms to keep time to a precision of better than a microsecond per year. The fundamental unit of time, the second, is based on such clocks. This image is looking down from the top of an atomic fountain nearly 30 feet tall! (credit: Steve Jurvetson/Flickr)

The Meter

The SI unit for length is the **meter** (abbreviated m); its definition has also changed over time to become more accurate and precise. The meter was first defined in 1791 as 1/10,000,000 of the distance from the equator to the North Pole. This measurement was improved in 1889 by redefining the meter to be the distance between two engraved lines on a platinum-iridium bar now kept near Paris. By 1960, it had become possible to define the meter even more accurately in terms of the wavelength of light, so it was again redefined as 1,650,763.73 wavelengths of orange light emitted by krypton atoms. In 1983, the meter was given its present definition (partly for greater accuracy) as the distance light travels in a vacuum in 1/299,792,458 of a second. (See [link].) This change defines the speed of light to be exactly 299,792,458 meters per second. The length of the meter will change if the speed of light is someday measured with greater accuracy.

The Kilogram

The SI unit for mass is the **kilogram** (abbreviated kg); it is defined to be the mass of a platinum-iridium cylinder kept with the old meter standard at the International Bureau of Weights and Measures near Paris. Exact replicas of the standard kilogram are also kept at the United States' National Institute of Standards and Technology, or NIST, located in Gaithersburg, Maryland outside of Washington D.C., and at other locations around the world. The determination of all other masses can be ultimately traced to a comparison with the standard mass.



The meter is defined to be the distance light travels in 1/299,792,458 of a second in a vacuum. Distance traveled is speed multiplied by time.

Electric current and its accompanying unit, the ampere, will be introduced in Introduction to Electric Current, Resistance, and Ohm's Law when electricity and magnetism are covered. The initial modules in this textbook are concerned with mechanics, fluids, heat, and waves. In these subjects all pertinent physical quantities can be expressed in terms of the fundamental units of length, mass, and time.

Metric Prefixes

SI units are part of the **metric system**. The metric system is convenient for scientific and engineering calculations because the units are categorized by factors of 10. [link] gives metric prefixes and symbols used to denote various factors of 10.

Metric systems have the advantage that conversions of units involve only powers of 10. There are 100 centimeters in a meter, 1000 meters in a kilometer, and so on. In nonmetric systems, such as the system of U.S. customary units, the relationships are not as simple—there are 12 inches in a foot, 5280 feet in a mile, and so on. Another advantage of the metric system is that the same unit can be used over extremely large ranges of values simply by using an appropriate metric prefix. For example, distances in meters are suitable in construction, while distances in kilometers are appropriate for air travel, and the tiny measure of nanometers are convenient in optical design. With the metric system there is no need to invent new units for particular applications.

The term **order of magnitude** refers to the scale of a value expressed in the metric system. Each power of 10 in the metric system represents a different order of magnitude. For example, 10^1, 10^2, 10^3, and so forth are all different orders of magnitude. All quantities that can be expressed as a product of a specific power of 10 are said to be of the *same* order of magnitude. For example, the number 800 can be written as 8×10^2, and the number 450 can be written as 4.5×10^2. Thus, the numbers 800 and 450 are of the same order of magnitude: 10^2. Order of magnitude can be thought of as a ballpark estimate for the scale of a value. The diameter of an atom is on the order of $10^{-9}\ \text{m}$, while the diameter of the Sun is on the order of $10^9\ \text{m}$.

Note:
The Quest for Microscopic Standards for Basic Units

The fundamental units described in this chapter are those that produce the greatest accuracy and precision in measurement. There is a sense among physicists that, because there is an underlying microscopic substructure to matter, it would be most satisfying to base our standards of measurement on microscopic objects and fundamental physical phenomena such as the speed of light. A microscopic standard has been accomplished for the standard of time, which is based on the oscillations of the cesium atom.

The standard for length was once based on the wavelength of light (a small-scale length) emitted by a certain type of atom, but it has been supplanted by the more precise measurement of the speed of light. If it becomes possible to measure the mass of atoms or a particular arrangement of atoms such as a silicon sphere to greater precision than the kilogram standard, it may become possible to base mass measurements on the small scale. There are also possibilities that electrical phenomena on the small scale may someday allow us to base a unit of charge on the charge of electrons and protons, but at present current and charge are related to large-scale currents and forces between wires.

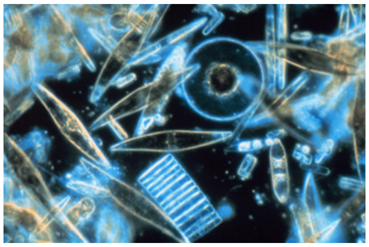
Prefix	**Symbol**	**Value[footnote] See Appendix A for a discussion of powers of 10.**	**Example (some are approximate)**			
exa	E	10^{18}	exameter	Em	10^{18} m	distance light travels in a century
peta	P	10^{15}	petasecond	Ps	10^{15} s	30 million years
tera	T	10^{12}	terawatt	TW	10^{12} W	powerful laser output
giga	G	10^{9}	gigahertz	GHz	10^{9} Hz	a microwave frequency
mega	M	10^{6}	megacurie	MCi	10^{6} Ci	high radioactivity
kilo	k	10^{3}	kilometer	km	10^{3} m	about 6/10 mile
hecto	h	10^{2}	hectoliter	hL	10^{2} L	26 gallons

Prefix	Symbol	**Value[footnote] See Appendix A for a discussion of powers of 10.**	**Example (some are approximate)**			
deka	da	10^1	dekagram	dag	10^1 g	teaspoon of butter
—	—	10^0 (=1)				
deci	d	10^{-1}	deciliter	dL	10^{-1} L	less than half a soda
centi	c	10^{-2}	centimeter	cm	10^{-2} m	fingertip thickness
milli	m	10^{-3}	millimeter	mm	10^{-3} m	flea at its shoulders
micro	µ	10^{-6}	micrometer	µm	10^{-6} m	detail in microscope
nano	n	10^{-9}	nanogram	ng	10^{-9} g	small speck of dust
pico	p	10^{-12}	picofarad	pF	10^{-12} F	small capacitor in radio
femto	f	10^{-15}	femtometer	fm	10^{-15} m	size of a proton
atto	a	10^{-18}	attosecond	as	10^{-18} s	time light crosses an atom

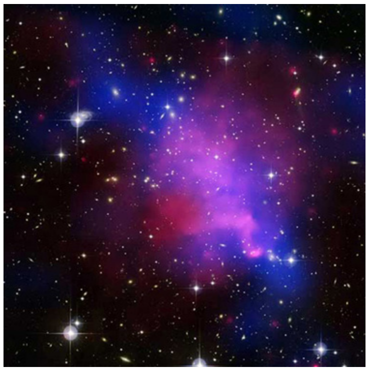
Metric Prefixes for Powers of 10 and their Symbols

Known Ranges of Length, Mass, and Time

The vastness of the universe and the breadth over which physics applies are illustrated by the wide range of examples of known lengths, masses, and times in [link]. Examination of this table will give you some feeling for the range of possible topics and numerical values. (See [link] and [link].)



Tiny phytoplankton swims among crystals of ice in the Antarctic Sea. They range from a few micrometers to as much as 2 millimeters in length. (credit: Prof. Gordon T. Taylor, Stony Brook University; NOAA Corps Collections)



Galaxies collide 2.4 billion light years away from Earth. The tremendous range of observable phenomena in nature challenges the imagination. (credit: NASA/CXC/UVic./A. Mahdavi et al. Optical/lensing: CFHT/UVic./H. Hoekstra et al.)

Unit Conversion and Dimensional Analysis

It is often necessary to convert from one type of unit to another. For example, if you are reading a European cookbook, some quantities may be expressed in units of liters and you need to convert them to cups. Or, perhaps you are reading walking directions from one location to another and you are interested in how many miles you will be walking. In this case, you will need to convert units of feet to miles.

Let us consider a simple example of how to convert units. Let us say that we want to convert 80 meters (m) to kilometers (km).

The first thing to do is to list the units that you have and the units that you want to convert to. In this case, we have units in *meters* and we want to convert to *kilometers*.

Next, we need to determine a **conversion factor** relating meters to kilometers. A conversion factor is a ratio expressing how many of one unit are equal to another unit. For example, there are 12 inches in 1 foot, 100 centimeters in 1 meter, 60 seconds in 1 minute, and so on. In this case, we know that there are 1,000 meters in 1 kilometer.

Now we can set up our unit conversion. We will write the units that we have and then multiply them by the conversion factor so that the units cancel out, as shown:
Equation:

$$80\,\cancel{\text{m}} \times \frac{1\text{ km}}{1000\,\cancel{\text{m}}} = 0.080\text{ km}.$$

Note that the unwanted m unit cancels, leaving only the desired km unit. You can use this method to convert between any types of unit.

Click [link] for a more complete list of conversion factors.

Lengths in meters		**Masses in kilograms (more precise values in parentheses)**		**Times in seconds (more precise values in parentheses)**	
10^{-18}	Present experimental limit to smallest observable detail	10^{-30}	Mass of an electron $\left(9.11\times10^{-31}\text{ kg}\right)$	10^{-23}	Time for light to cross a proton
10^{-15}	Diameter of a proton	10^{-27}	Mass of a hydrogen atom $\left(1.67\times10^{-27}\text{ kg}\right)$	10^{-22}	Mean life of an extremely unstable nucleus

Lengths in meters		Masses in kilograms (more precise values in parentheses)		Times in seconds (more precise values in parentheses)	
10^{-14}	Diameter of a uranium nucleus	10^{-15}	Mass of a bacterium	10^{-15}	Time for one oscillation of visible light
10^{-10}	Diameter of a hydrogen atom	10^{-5}	Mass of a mosquito	10^{-13}	Time for one vibration of an atom in a solid
10^{-8}	Thickness of membranes in cells of living organisms	10^{-2}	Mass of a hummingbird	10^{-8}	Time for one oscillation of an FM radio wave
10^{-6}	Wavelength of visible light	1	Mass of a liter of water (about a quart)	10^{-3}	Duration of a nerve impulse
10^{-3}	Size of a grain of sand	10^{2}	Mass of a person	1	Time for one heartbeat
1	Height of a 4-year-old child	10^{3}	Mass of a car	10^{5}	One day $\left(8.64\times10^{4}\text{ s}\right)$
10^{2}	Length of a football field	10^{8}	Mass of a large ship	10^{7}	One year (y) $\left(3.16\times10^{7}\text{ s}\right)$
10^{4}	Greatest ocean depth	10^{12}	Mass of a large iceberg	10^{9}	About half the life expectancy of a human
10^{7}	Diameter of the Earth	10^{15}	Mass of the nucleus of a comet	10^{11}	Recorded history
10^{11}	Distance from the Earth to the Sun	10^{23}	Mass of the Moon $\left(7.35\times10^{22}\text{ kg}\right)$	10^{17}	Age of the Earth
10^{16}	Distance traveled by light in 1 year (a light year)	10^{25}	Mass of the Earth $\left(5.97\times10^{24}\text{ kg}\right)$	10^{18}	Age of the universe
10^{21}	Diameter of the Milky Way galaxy	10^{30}	Mass of the Sun $\left(1.99\times10^{30}\text{ kg}\right)$		

Lengths in meters		**Masses in kilograms (more precise values in parentheses)**		**Times in seconds (more precise values in parentheses)**	
10^{22}	Distance from the Earth to the nearest large galaxy (Andromeda)	10^{42}	Mass of the Milky Way galaxy (current upper limit)		
10^{26}	Distance from the Earth to the edges of the known universe	10^{53}	Mass of the known universe (current upper limit)		

Approximate Values of Length, Mass, and Time

Example:
Unit Conversions: A Short Drive Home

Suppose that you drive the 10.0 km from your university to home in 20.0 min. Calculate your average speed (a) in kilometers per hour (km/h) and (b) in meters per second (m/s). (Note: Average speed is distance traveled divided by time of travel.)

Strategy

First we calculate the average speed using the given units. Then we can get the average speed into the desired units by picking the correct conversion factor and multiplying by it. The correct conversion factor is the one that cancels the unwanted unit and leaves the desired unit in its place.

Solution for (a)

(1) Calculate average speed. Average speed is distance traveled divided by time of travel. (Take this definition as a given for now—average speed and other motion concepts will be covered in a later module.) In equation form,

Equation:

$$\text{average speed} = \frac{\text{distance}}{\text{time}}.$$

(2) Substitute the given values for distance and time.

Equation:

$$\text{average speed} = \frac{10.0\text{ km}}{20.0\text{ min}} = 0.500\frac{\text{km}}{\text{min}}.$$

(3) Convert km/min to km/h: multiply by the conversion factor that will cancel minutes and leave hours. That conversion factor is $60\text{ min}/\text{hr}$. Thus,

Equation:

$$\text{average speed} = 0.500\frac{\text{km}}{\text{min}} \times \frac{60\text{ min}}{1\text{ h}} = 30.0\frac{\text{km}}{\text{h}}.$$

Discussion for (a)

To check your answer, consider the following:

(1) Be sure that you have properly cancelled the units in the unit conversion. If you have written the unit conversion factor upside down, the units will not cancel properly in the equation. If you accidentally get the ratio upside down, then the units will not cancel; rather, they will give you the wrong units as follows:

Equation:

$$\frac{\text{km}}{\text{min}} \times \frac{1\ \text{hr}}{60\ \text{min}} = \frac{1}{60}\ \frac{\text{km} \cdot \text{hr}}{\text{min}^2},$$

which are obviously not the desired units of km/h.

(2) Check that the units of the final answer are the desired units. The problem asked us to solve for average speed in units of km/h and we have indeed obtained these units.

(3) Check the significant figures. Because each of the values given in the problem has three significant figures, the answer should also have three significant figures. The answer 30.0 km/hr does indeed have three significant figures, so this is appropriate. Note that the significant figures in the conversion factor are not relevant because an hour is *defined* to be 60 minutes, so the precision of the conversion factor is perfect.

(4) Next, check whether the answer is reasonable. Let us consider some information from the problem—if you travel 10 km in a third of an hour (20 min), you would travel three times that far in an hour. The answer does seem reasonable.

Solution for (b)

There are several ways to convert the average speed into meters per second.

(1) Start with the answer to (a) and convert km/h to m/s. Two conversion factors are needed—one to convert hours to seconds, and another to convert kilometers to meters.

(2) Multiplying by these yields

Equation:

$$\text{Average speed} = 30.0\frac{\text{km}}{\text{h}} \times \frac{1\ \text{h}}{3{,}600\ \text{s}} \times \frac{1{,}000\ \text{m}}{1\ \text{km}},$$

Equation:

$$\text{Average speed} = 8.33\frac{\text{m}}{\text{s}}.$$

Discussion for (b)

If we had started with 0.500 km/min, we would have needed different conversion factors, but the answer would have been the same: 8.33 m/s.

You may have noted that the answers in the worked example just covered were given to three digits. Why? When do you need to be concerned about the number of digits in something you calculate? Why not write down all the digits your calculator produces? The module Accuracy, Precision, and Significant Figures will help you answer these questions.

Note:

Nonstandard Units

While there are numerous types of units that we are all familiar with, there are others that are much more obscure. For example, a **firkin** is a unit of volume that was once used to measure beer. One firkin equals about 34 liters. To learn more about nonstandard units, use a dictionary or encyclopedia to research different “weights and measures.” Take note of any unusual units, such as a barleycorn, that are not listed in the text. Think about how the unit is defined and state its relationship to SI units.

Exercise:
Check Your Understanding

Problem:

Some hummingbirds beat their wings more than 50 times per second. A scientist is measuring the time it takes for a hummingbird to beat its wings once. Which fundamental unit should the scientist use to describe the measurement? Which factor of 10 is the scientist likely to use to describe the motion precisely? Identify the metric prefix that corresponds to this factor of 10.

Solution:

The scientist will measure the time between each movement using the fundamental unit of seconds. Because the wings beat so fast, the scientist will probably need to measure in milliseconds, or 10^{-3} seconds. (50 beats per second corresponds to 20 milliseconds per beat.)

Exercise:
Check Your Understanding

Problem:

One cubic centimeter is equal to one milliliter. What does this tell you about the different units in the SI metric system?

Solution:

The fundamental unit of length (meter) is probably used to create the derived unit of volume (liter). The measure of a milliliter is dependent on the measure of a centimeter.

Summary

- Physical quantities are a characteristic or property of an object that can be measured or calculated from other measurements.
- Units are standards for expressing and comparing the measurement of physical quantities. All units can be expressed as combinations of four fundamental units.
- The four fundamental units we will use in this text are the meter (for length), the kilogram (for mass), the second (for time), and the ampere (for electric current). These units are part of the metric system, which uses powers of 10 to relate quantities over the vast ranges encountered in nature.
- The four fundamental units are abbreviated as follows: meter, m; kilogram, kg; second, s; and ampere, A. The metric system also uses a standard set of prefixes to denote each order of magnitude greater than or lesser than the fundamental unit itself.
- Unit conversions involve changing a value expressed in one type of unit to another type of unit. This is done by using conversion factors, which are ratios relating equal quantities of different units.

Conceptual Questions

Exercise:

Problem: Identify some advantages of metric units.

Problems & Exercises

Exercise:

Problem:

The speed limit on some interstate highways is roughly 100 km/h. (a) What is this in meters per second? (b) How many miles per hour is this?

Solution:

a. $27.8\ \text{m/s}$
b. $62.1\ \text{mph}$

Exercise:

Problem:

A car is traveling at a speed of $33\ \text{m/s}$. (a) What is its speed in kilometers per hour? (b) Is it exceeding the $90\ \text{km/h}$ speed limit?

Exercise:

Problem:

Show that $1.0\ \text{m/s} = 3.6\ \text{km/h}$. Hint: Show the explicit steps involved in converting $1.0\ \text{m/s} = 3.6\ \text{km/h}$.

Solution:

$\frac{1.0\ \text{m}}{\text{s}} = \frac{1.0\ \text{m}}{\text{s}} \times \frac{3600\ \text{s}}{1\ \text{hr}} \times \frac{1\ \text{km}}{1000\ \text{m}}$

$= 3.6\ \text{km/h}.$

Exercise:

Problem:

American football is played on a 100-yd-long field, excluding the end zones. How long is the field in meters? (Assume that 1 meter equals 3.281 feet.)

Exercise:

Problem:

Soccer fields vary in size. A large soccer field is 115 m long and 85 m wide. What are its dimensions in feet and inches? (Assume that 1 meter equals 3.281 feet.)

Solution:

length: $377\ \text{ft}$; $4.53 \times 10^3\ \text{in}$. width: $280\ \text{ft}$; $3.3 \times 10^3\ \text{in}$.

Exercise:

Problem:

What is the height in meters of a person who is 6 ft 1.0 in. tall? (Assume that 1 meter equals 39.37 in.)

Exercise:

Problem:

Mount Everest, at 29,028 feet, is the tallest mountain on the Earth. What is its height in kilometers? (Assume that 1 kilometer equals 3,281 feet.)

Solution:

$8.847\ \text{km}$

Exercise:

Problem: The speed of sound is measured to be $342\ \text{m/s}$ on a certain day. What is this in km/h?

Exercise:

Problem:

Tectonic plates are large segments of the Earth’s crust that move slowly. Suppose that one such plate has an average speed of 4.0 cm/year. (a) What distance does it move in 1 s at this speed? (b) What is its speed in kilometers per million years?

Solution:

(a) $1.3 \times 10^{-9}\ \text{m}$

(b) $40\ \text{km/My}$

Exercise:

Problem:

(a) Refer to [link] to determine the average distance between the Earth and the Sun. Then calculate the average speed of the Earth in its orbit in kilometers per second. (b) What is this in meters per second?

Glossary

physical quantity
: a characteristic or property of an object that can be measured or calculated from other measurements

units
: a standard used for expressing and comparing measurements

SI units
: the international system of units that scientists in most countries have agreed to use; includes units such as meters, liters, and grams

English units
: system of measurement used in the United States; includes units of measurement such as feet, gallons, and pounds

fundamental units
: units that can only be expressed relative to the procedure used to measure them

derived units

units that can be calculated using algebraic combinations of the fundamental units

second

the SI unit for time, abbreviated (s)

meter

the SI unit for length, abbreviated (m)

kilogram

the SI unit for mass, abbreviated (kg)

metric system

a system in which values can be calculated in factors of 10

order of magnitude

refers to the size of a quantity as it relates to a power of 10

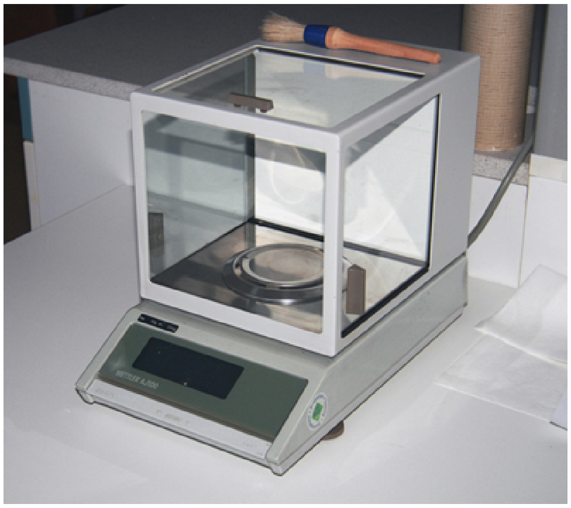
conversion factor

a ratio expressing how many of one unit are equal to another unit

Accuracy, Precision, and Significant Figures

- Determine the appropriate number of significant figures in both addition and subtraction, as well as multiplication and division calculations.
- Calculate the percent uncertainty of a measurement.

A double-pan mechanical balance is used to compare different masses. Usually an object with unknown mass is placed in one pan and objects of known mass are placed in the other pan. When the bar that connects the two pans is horizontal, then the masses in both pans are equal. The “known masses” are typically metal cylinders of standard mass such as 1 gram, 10 grams, and 100 grams. (credit: Serge Melki)



Many mechanical balances, such as double-pan balances, have been replaced by digital scales, which can typically measure the mass of an object more precisely. Whereas a mechanical balance may only read the mass of an object to the nearest tenth of a gram, many digital scales can measure the mass of an object up to the nearest thousandth of a gram. (credit: Karel Jakubec)

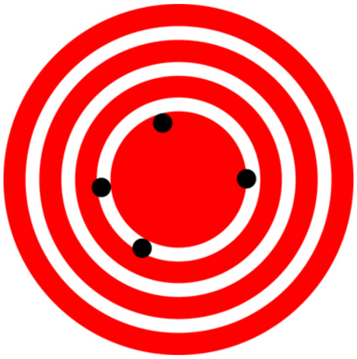
Accuracy and Precision of a Measurement

Science is based on observation and experiment—that is, on measurements. **Accuracy** is how close a measurement is to the correct value for that measurement. For example, let us say that you are measuring the length of standard computer paper. The packaging in which you purchased the paper states that it is 11.0 inches long. You measure the length of the paper three times and obtain the following measurements: 11.1 in., 11.2 in., and 10.9 in.

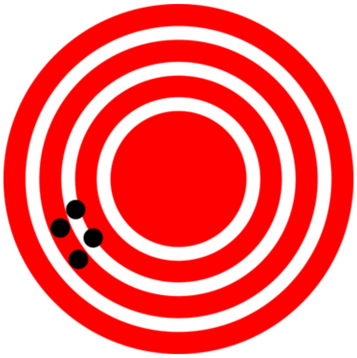
These measurements are quite accurate because they are very close to the correct value of 11.0 inches. In contrast, if you had obtained a measurement of 12 inches, your measurement would not be very accurate.

The **precision** of a measurement system is refers to how close the agreement is between repeated measurements (which are repeated under the same conditions). Consider the example of the paper measurements. The precision of the measurements refers to the spread of the measured values. One way to analyze the precision of the measurements would be to determine the range, or difference, between the lowest and the highest measured values. In that case, the lowest value was 10.9 in. and the highest value was 11.2 in. Thus, the measured values deviated from each other by at most 0.3 in. These measurements were relatively precise because they did not vary too much in value. However, if the measured values had been 10.9, 11.1, and 11.9, then the measurements would not be very precise because there would be significant variation from one measurement to another.

The measurements in the paper example are both accurate and precise, but in some cases, measurements are accurate but not precise, or they are precise but not accurate. Let us consider an example of a GPS system that is attempting to locate the position of a restaurant in a city. Think of the restaurant location as existing at the center of a bull’s-eye target, and think of each GPS attempt to locate the restaurant as a black dot. In [link], you can see that the GPS measurements are spread out far apart from each other, but they are all relatively close to the actual location of the restaurant at the center of the target. This indicates a low precision, high accuracy measuring system. However, in [link], the GPS measurements are concentrated quite closely to one another, but they are far away from the target location. This indicates a high precision, low accuracy measuring system.



A GPS system attempts to locate a restaurant at the center of the bull’s-eye. The black dots represent each attempt to pinpoint the location of the restaurant. The dots are spread out quite far apart from one another, indicating low precision, but they are each rather close to the actual location of the restaurant, indicating high accuracy. (credit: Dark Evil)



In this figure, the dots are concentrated rather closely to one another, indicating high precision, but they are rather far away from the actual location of the restaurant, indicating low accuracy. (credit: Dark Evil)

Accuracy, Precision, and Uncertainty

The degree of accuracy and precision of a measuring system are related to the **uncertainty** in the measurements. Uncertainty is a quantitative measure of how much your measured values deviate from a standard or expected value. If your measurements are not very accurate or precise, then the uncertainty of your values will be very high. In more general terms, uncertainty can be thought of as a disclaimer for your measured values. For example, if someone asked you to provide the mileage on your car, you might say that it is 45,000 miles, plus or minus 500 miles. The plus or minus amount is the uncertainty in your value. That is, you are indicating that the actual mileage of your car might be as low as 44,500 miles or as high as 45,500 miles, or anywhere in between. All measurements contain some amount of uncertainty. In our example of measuring the length of the paper, we might say that the length of the paper is 11 in., plus or minus 0.2 in. The uncertainty in a measurement, A, is often denoted as δA ("delta A "), so the measurement result would be recorded as $A \pm \delta A$. In our paper example, the length of the paper could be expressed as $11 \text{ in.} \pm 0.2.$

The factors contributing to uncertainty in a measurement include:

1. Limitations of the measuring device,
2. The skill of the person making the measurement,
3. Irregularities in the object being measured,
4. Any other factors that affect the outcome (highly dependent on the situation).

In our example, such factors contributing to the uncertainty could be the following: the smallest division on the ruler is 0.1 in., the person using the ruler has bad eyesight, or one side of the paper is slightly longer than the other. At any rate, the uncertainty in a measurement must be based on a careful consideration of all the factors that might contribute and their possible effects.

Note:
Making Connections: Real-World Connections – Fevers or Chills?
Uncertainty is a critical piece of information, both in physics and in many other real-world applications. Imagine you are caring for a sick child. You suspect the child has a fever, so you check his or her temperature with a thermometer. What if the uncertainty of the thermometer were 3.0°C? If the child's temperature reading was 37.0°C (which is normal body temperature), the "true" temperature could be anywhere from a

hypothermic 34.0^{o}C to a dangerously high 40.0^{o}C. A thermometer with an uncertainty of 3.0^{o}C would be useless.

Percent Uncertainty

One method of expressing uncertainty is as a percent of the measured value. If a measurement A is expressed with uncertainty, δA, the **percent uncertainty** (%unc) is defined to be

Equation:

$$\%\text{ unc} = \frac{\delta A}{A} \times 100\%.$$

Example:

Calculating Percent Uncertainty: A Bag of Apples

A grocery store sells 5-lb bags of apples. You purchase four bags over the course of a month and weigh the apples each time. You obtain the following measurements:

Week 1 weight: 4.8 lb
Week 2 weight: 5.3 lb
Week 3 weight: 4.9 lb
Week 4 weight: 5.4 lb

You determine that the weight of the 5-lb bag has an uncertainty of ±0.4 lb. What is the percent uncertainty of the bag's weight?

Strategy

First, observe that the expected value of the bag's weight, A, is 5 lb. The uncertainty in this value, δA, is 0.4 lb. We can use the following equation to determine the percent uncertainty of the weight:

Equation:

$$\% \text{ unc} = \frac{\delta A}{A} \times 100\%.$$

Solution
Plug the known values into the equation:
Equation:

$$\% \text{ unc} = \frac{0.4 \text{ lb}}{5 \text{ lb}} \times 100\% = 8\%.$$

Discussion
We can conclude that the weight of the apple bag is $5 \text{ lb} \pm 8\%$. Consider how this percent uncertainty would change if the bag of apples were half as heavy, but the uncertainty in the weight remained the same. Hint for future calculations: when calculating percent uncertainty, always remember that you must multiply the fraction by 100%. If you do not do this, you will have a decimal quantity, not a percent value.

Uncertainties in Calculations

There is an uncertainty in anything calculated from measured quantities. For example, the area of a floor calculated from measurements of its length and width has an uncertainty because the length and width have uncertainties. How big is the uncertainty in something you calculate by multiplication or division? If the measurements going into the calculation have small uncertainties (a few percent or less), then the **method of adding percents** can be used for multiplication or division. This method says that *the percent uncertainty in a quantity calculated by multiplication or division is the sum of the percent uncertainties in the items used to make the calculation*. For example, if a floor has a length of 4.00 m and a width of 3.00 m, with uncertainties of 2% and 1%, respectively, then the area of the floor is 12.0 m^2 and has an uncertainty of 3%. (Expressed as an area this is 0.36 m^2, which we round to 0.4 m^2 since the area of the floor is given to a tenth of a square meter.)
Exercise:
Check Your Understanding

Problem:

A high school track coach has just purchased a new stopwatch. The stopwatch manual states that the stopwatch has an uncertainty of ± 0.05 s. Runners on the track coach's team regularly clock 100-m sprints of 11.49 s to 15.01 s. At the school's last track meet, the first-place sprinter came in at 12.04 s and the second-place sprinter came in at 12.07 s. Will the coach's new stopwatch be helpful in timing the sprint team? Why or why not?

Solution:

No, the uncertainty in the stopwatch is too great to effectively differentiate between the sprint times.

Precision of Measuring Tools and Significant Figures

An important factor in the accuracy and precision of measurements involves the precision of the measuring tool. In general, a precise measuring tool is one that can measure values in very small increments. For example, a standard ruler can measure length to the nearest millimeter, while a caliper can measure length to the nearest 0.01 millimeter. The caliper is a more precise measuring tool because it can measure extremely small differences in length. The more precise the measuring tool, the more precise and accurate the measurements can be.

When we express measured values, we can only list as many digits as we initially measured with our measuring tool. For example, if you use a standard ruler to measure the length of a stick, you may measure it to be 36.7 cm. You could not express this value as 36.71 cm because your measuring tool was not precise enough to measure a hundredth of a centimeter. It should be noted that the last digit in a measured value has been estimated in some way by the person performing the measurement. For example, the person measuring the length of a stick with a ruler notices that the stick length seems to be somewhere in between 36.6 cm and 36.7 cm, and he or she must estimate the value of the last digit. Using the method of **significant figures**, the rule is that *the last digit written down in a measurement is the first digit with some uncertainty*. In order to determine the number of significant digits in a value, start with the first measured value at the left and count the number of digits through the last digit written on the right. For example, the measured value 36.7 cm has three digits, or significant figures. Significant figures indicate the precision of a measuring tool that was used to measure a value.

Zeros

Special consideration is given to zeros when counting significant figures. The zeros in 0.053 are not significant, because they are only placekeepers that locate the decimal point. There are two significant figures in 0.053. The zeros in 10.053 are not placekeepers but are significant—this number has five significant figures. The zeros in 1300 may or may not be significant depending on the style of writing numbers. They could mean the number is known to the last digit, or they could be placekeepers. So 1300 could have two, three, or four significant figures. (To avoid this ambiguity, write 1300 in scientific notation.) *Zeros are significant except when they serve only as placekeepers*.

Exercise:

Check Your Understanding

Problem:

Determine the number of significant figures in the following measurements:

a. 0.0009
b. 15,450.0
c. 6×10^3
d. 87.990
e. 30.42

Solution:

(a) 1; the zeros in this number are placekeepers that indicate the decimal point

(b) 6; here, the zeros indicate that a measurement was made to the 0.1 decimal point, so the zeros are significant

(c) 1; the value 10^3 signifies the decimal place, not the number of measured values

(d) 5; the final zero indicates that a measurement was made to the 0.001 decimal point, so it is significant

(e) 4; any zeros located in between significant figures in a number are also significant

Significant Figures in Calculations

When combining measurements with different degrees of accuracy and precision, *the number of significant digits in the final answer can be no greater than the number of significant digits in the least precise measured value*. There are two different rules, one for multiplication and division and the other for addition and subtraction, as discussed below.

1. For multiplication and division: *The result should have the same number of significant figures as the quantity having the least significant figures entering into the calculation*. For example, the area of a circle can be calculated from its radius using $A = \pi r^2$. Let us see how many significant figures the area has if the radius has only two—say, $r = 1.2\text{ m}$. Then,

Equation:

$$A = \pi r^2 = (3.1415927...) \times (1.2\text{ m})^2 = 4.5238934\text{ m}^2$$

is what you would get using a calculator that has an eight-digit output. But because the radius has only two significant figures, it limits the calculated

quantity to two significant figures or
Equation:

$$A=4.5\ \text{m}^2,$$

even though π is good to at least eight digits.

2. For addition and subtraction: *The answer can contain no more decimal places than the least precise measurement.* Suppose that you buy 7.56-kg of potatoes in a grocery store as measured with a scale with precision 0.01 kg. Then you drop off 6.052-kg of potatoes at your laboratory as measured by a scale with precision 0.001 kg. Finally, you go home and add 13.7 kg of potatoes as measured by a bathroom scale with precision 0.1 kg. How many kilograms of potatoes do you now have, and how many significant figures are appropriate in the answer? The mass is found by simple addition and subtraction:
Equation:

$$\begin{array}{ll} & 7.56\ \ \ \text{kg} \\ - & 6.052\ \text{kg} \\ & \frac{+13.7\ \ \ \ \text{kg}}{15.208\ \text{kg}} = 15.2\ \text{kg}. \end{array}$$

Next, we identify the least precise measurement: 13.7 kg. This measurement is expressed to the 0.1 decimal place, so our final answer must also be expressed to the 0.1 decimal place. Thus, the answer is rounded to the tenths place, giving us 15.2 kg.

Significant Figures in this Text

In this text, most numbers are assumed to have three significant figures. Furthermore, consistent numbers of significant figures are used in all worked examples. You will note that an answer given to three digits is based on input good to at least three digits, for example. If the input has fewer significant figures, the answer will also have fewer significant figures. Care is also taken that the number of significant figures is reasonable for the situation posed. In some topics, particularly in optics, more accurate numbers are needed and more than three significant figures will be used. Finally, if a number is *exact*, such as the two in the formula for the circumference of a circle, $c = 2\pi r$, it does not affect the number of significant figures in a calculation.

Exercise:

Check Your Understanding

Problem:

Perform the following calculations and express your answer using the correct number of significant digits.

(a) A woman has two bags weighing 13.5 pounds and one bag with a weight of 10.2 pounds. What is the total weight of the bags?

(b) The force F on an object is equal to its mass m multiplied by its acceleration a. If a wagon with mass 55 kg accelerates at a rate of $0.0255\ \text{m/s}^2$, what is the force on the wagon? (The unit of force is called the newton, and it is expressed with the symbol N.)

Solution:

(a) 37.2 pounds; Because the number of bags is an exact value, it is not considered in the significant figures.

(b) 1.4 N; Because the value 55 kg has only two significant figures, the final value must also contain two significant figures.

Note:

PhET Explorations: Estimation

Explore size estimation in one, two, and three dimensions! Multiple levels of difficulty allow for progressive skill improvement.

https://phet.colorado.edu/sims/estimation/estimation_en.html

Summary

- Accuracy of a measured value refers to how close a measurement is to the correct value. The uncertainty in a measurement is an estimate of the amount by which the measurement result may differ from this value.
- Precision of measured values refers to how close the agreement is between repeated measurements.
- The precision of a *measuring tool* is related to the size of its measurement increments. The smaller the measurement increment, the more precise the tool.
- Significant figures express the precision of a measuring tool.
- When multiplying or dividing measured values, the final answer can contain only as many significant figures as the least precise value.
- When adding or subtracting measured values, the final answer cannot contain more decimal places than the least precise value.

Conceptual Questions

Exercise:

Problem:

What is the relationship between the accuracy and uncertainty of a measurement?

Exercise:

Problem:

Prescriptions for vision correction are given in units called *diopters* (D). Determine the meaning of that unit. Obtain information (perhaps by calling an optometrist or performing an internet search) on the minimum uncertainty with which corrections in diopters are determined and the accuracy with which corrective lenses can be produced. Discuss the sources of uncertainties in both the prescription and accuracy in the manufacture of lenses.

Problems & Exercises

Express your answers to problems in this section to the correct number of significant figures and proper units.

Exercise:

Problem:

Suppose that your bathroom scale reads your mass as 65 kg with a 3% uncertainty. What is the uncertainty in your mass (in kilograms)?

Solution:

2 kg

Exercise:

Problem:

A good-quality measuring tape can be off by 0.50 cm over a distance of 20 m. What is its percent uncertainty?

Exercise:

Problem:

(a) A car speedometer has a 5.0% uncertainty. What is the range of possible speeds when it reads $90\ \text{km/h}$? (b) Convert this range to miles per hour. $(1\ \text{km} = 0.6214\ \text{mi})$

Solution:

a. $85.5\text{ to }94.5\ \text{km/h}$
b. $53.1\text{ to }58.7\ \text{mi/h}$

Exercise:

Problem:

An infant’s pulse rate is measured to be 130 ± 5 beats/min. What is the percent uncertainty in this measurement?

Exercise:

Problem:

(a) Suppose that a person has an average heart rate of 72.0 beats/min. How many beats does he or she have in 2.0 y? (b) In 2.00 y? (c) In 2.000 y?

Solution:

(a) 7.6×10^7 beats

(b) 7.57×10^7 beats

(c) 7.57×10^7 beats

Exercise:

Problem:

A can contains 375 mL of soda. How much is left after 308 mL is removed?

Exercise:

Problem:

State how many significant figures are proper in the results of the following calculations: (a) $(106.7)(98.2)/(46.210)(1.01)$ (b) $(18.7)^2$ (c) $\left(1.60 \times 10^{-19}\right)(3712)$.

Solution:

a. 3
b. 3
c. 3

Exercise:

Problem:

(a) How many significant figures are in the numbers 99 and 100? (b) If the uncertainty in each number is 1, what is the percent uncertainty in each? (c) Which is a more meaningful way to express the accuracy of these two numbers, significant figures or percent uncertainties?

Exercise:

Problem:

(a) If your speedometer has an uncertainty of 2.0 km/h at a speed of 90 km/h, what is the percent uncertainty? (b) If it has the same percent uncertainty when it reads 60 km/h, what is the range of speeds you could be going?

Solution:

a) 2.2%

(b) $59\text{ to }61\text{ km/h}$

Exercise:

Problem:

(a) A person's blood pressure is measured to be $120\pm 2\text{ mm Hg}$. What is its percent uncertainty? (b) Assuming the same percent uncertainty, what is the uncertainty in a blood pressure measurement of 80 mm Hg?

Exercise:

Problem:

A person measures his or her heart rate by counting the number of beats in 30 s. If 40 ± 1 beats are counted in $30.0\pm 0.5\text{ s}$, what is the heart rate and its uncertainty in beats per minute?

Solution:

$80 \pm 3 \text{ beats/min}$

Exercise:

Problem: What is the area of a circle 3.102 cm in diameter?

Exercise:

Problem:

If a marathon runner averages 9.5 mi/h, how long does it take him or her to run a 26.22-mi marathon?

Solution:

2.8 h

Exercise:

Problem:

A marathon runner completes a 42.188-km course in 2 h, 30 min, and 12 s. There is an uncertainty of 25 m in the distance traveled and an uncertainty of 1 s in the elapsed time. (a) Calculate the percent uncertainty in the distance. (b) Calculate the uncertainty in the elapsed time. (c) What is the average speed in meters per second? (d) What is the uncertainty in the average speed?

Exercise:

Problem:

The sides of a small rectangular box are measured to be $1.80 \pm 0.01 \text{ cm}$, $2.05 \pm 0.02 \text{ cm}$, and $3.1 \pm 0.1 \text{ cm}$ long. Calculate its volume and uncertainty in cubic centimeters.

Solution:

$11 \pm 1 \text{ cm}^3$

Exercise:

Problem:

When non-metric units were used in the United Kingdom, a unit of mass called the *pound-mass* (lbm) was employed, where $1\ \text{lbm} = 0.4539\ \text{kg}$. (a) If there is an uncertainty of $0.0001\ \text{kg}$ in the pound-mass unit, what is its percent uncertainty? (b) Based on that percent uncertainty, what mass in pound-mass has an uncertainty of 1 kg when converted to kilograms?

Exercise:

Problem:

The length and width of a rectangular room are measured to be $3.955 \pm 0.005\ \text{m}$ and $3.050 \pm 0.005\ \text{m}$. Calculate the area of the room and its uncertainty in square meters.

Solution:

$12.06 \pm 0.04\ \text{m}^2$

Exercise:

Problem:

A car engine moves a piston with a circular cross section of $7.500 \pm 0.002\ \text{cm}$ diameter a distance of $3.250 \pm 0.001\ \text{cm}$ to compress the gas in the cylinder. (a) By what amount is the gas decreased in volume in cubic centimeters? (b) Find the uncertainty in this volume.

Glossary

accuracy

the degree to which a measured value agrees with correct value for that measurement

method of adding percents

the percent uncertainty in a quantity calculated by multiplication or division is the sum of the percent uncertainties in the items used to make the calculation

percent uncertainty
the ratio of the uncertainty of a measurement to the measured value, expressed as a percentage

precision
the degree to which repeated measurements agree with each other

significant figures
express the precision of a measuring tool used to measure a value

uncertainty
a quantitative measure of how much your measured values deviate from a standard or expected value

Approximation

- Make reasonable approximations based on given data.

On many occasions, physicists, other scientists, and engineers need to make **approximations** or “guesstimates” for a particular quantity. What is the distance to a certain destination? What is the approximate density of a given item? About how large a current will there be in a circuit? Many approximate numbers are based on formulae in which the input quantities are known only to a limited accuracy. As you develop problem-solving skills (that can be applied to a variety of fields through a study of physics), you will also develop skills at approximating. You will develop these skills through thinking more quantitatively, and by being willing to take risks. As with any endeavor, experience helps, as well as familiarity with units. These approximations allow us to rule out certain scenarios or unrealistic numbers. Approximations also allow us to challenge others and guide us in our approaches to our scientific world. Let us do two examples to illustrate this concept.

Example:

Approximate the Height of a Building

Can you approximate the height of one of the buildings on your campus, or in your neighborhood? Let us make an approximation based upon the height of a person. In this example, we will calculate the height of a 39-story building.

Strategy

Think about the average height of an adult male. We can approximate the height of the building by scaling up from the height of a person.

Solution

Based on information in the example, we know there are 39 stories in the building. If we use the fact that the height of one story is approximately equal to about the length of two adult humans (each human is about 2-m tall), then we can estimate the total height of the building to be

Equation:

$$\frac{2\text{ m}}{1\text{ person}} \times \frac{2\text{ person}}{1\text{ story}} \times 39\text{ stories} = 156\text{ m}.$$

Discussion

You can use known quantities to determine an approximate measurement of unknown quantities. If your hand measures 10 cm across, how many hand lengths equal the width of your desk? What other measurements can you approximate besides length?

Example:

Approximating Vast Numbers: a Trillion Dollars

A bank stack contains one-hundred $100 bills, and is worth $10,000. How many bank stacks make up a trillion dollars? (credit: Andrew Magill)

The U.S. federal deficit in the 2008 fiscal year was a little greater than $10 trillion. Most of us do not have any concept of how much even one trillion actually is. Suppose that you were given a trillion dollars in $100 bills. If you made 100-bill stacks and used them to evenly cover a football field (between the end zones), make an approximation of how high the money pile would become. (We will use feet/inches rather than meters here because football fields are measured in yards.) One of your friends says 3 in., while another says 10 ft. What do you think?

Strategy

When you imagine the situation, you probably envision thousands of small stacks of 100 wrapped $100 bills, such as you might see in movies or at a bank. Since this is an easy-to-approximate quantity, let us start there. We can find the volume of a stack of 100 bills, find out how many stacks make up one trillion dollars, and then set this volume equal to the area of the football field multiplied by the unknown height.

Solution

(1) Calculate the volume of a stack of 100 bills. The dimensions of a single bill are approximately 3 in. by 6 in. A stack of 100 of these is about 0.5 in. thick. So the total volume of a stack of 100 bills is:

Equation:

$$\begin{aligned}\text{volume of stack} &= \text{length} \times \text{width} \times \text{height},\\ \text{volume of stack} &= 6\text{ in.} \times 3\text{ in.} \times 0.5\text{ in.},\\ \text{volume of stack} &= 9\text{ in.}^3.\end{aligned}$$

(2) Calculate the number of stacks. Note that a trillion dollars is equal to $\$1 \times 10^{12}$, and a stack of one-hundred $100 bills is equal to $\$10{,}000$, or $\$1 \times 10^4$. The number of stacks you will have is:

Equation:

$$\$1 \times 10^{12}(\text{a trillion dollars})/\ \$1 \times 10^4 \text{ per stack} = 1 \times 10^8 \text{ stacks.}$$

(3) Calculate the area of a football field in square inches. The area of a football field is $100\text{ yd}\times 50\text{ yd}$, which gives $5{,}000\text{ yd}^2$. Because we are working in inches, we need to convert square yards to square inches:

Equation:

$$\begin{gathered}\text{Area} = 5{,}000\text{ yd}^2 \times \frac{3\text{ ft}}{1\text{ yd}} \times \frac{3\text{ ft}}{1\text{ yd}} \times \frac{12\text{ in.}}{1\text{ ft}} \times \frac{12\text{ in.}}{1\text{ ft}} = 6{,}480{,}000\text{ in.}^2,\\ \text{Area} \approx 6 \times 10^6\text{ in.}^2.\end{gathered}$$

This conversion gives us $6 \times 10^6\text{ in.}^2$ for the area of the field. (Note that we are using only one significant figure in these calculations.)

(4) Calculate the total volume of the bills. The volume of all the $100-bill stacks is $9\text{ in.}^3/\text{stack}\times 10^8\text{ stacks} = 9\times 10^8\text{ in.}^3$.

(5) Calculate the height. To determine the height of the bills, use the equation:

Equation:

$$\begin{aligned}\text{volume of bills} &= \text{area of field}\times\text{height of money:}\\ \text{Height of money} &= \tfrac{\text{volume of bills}}{\text{area of field}},\\ \text{Height of money} &= \tfrac{9\times10^8\text{in.}^3}{6\times10^6\text{in.}^2} = 1.33\times10^2\text{in.},\\ \text{Height of money} &\approx 1\times10^2\text{in.} = 100\text{ in.}\end{aligned}$$

The height of the money will be about 100 in. high. Converting this value to feet gives

Equation:

$$100\text{ in.}\times\frac{1\text{ ft}}{12\text{ in.}} = 8.33\text{ ft}\approx 8\text{ ft.}$$

Discussion

The final approximate value is much higher than the early estimate of 3 in., but the other early estimate of 10 ft (120 in.) was roughly correct. How did the approximation measure up to your first guess? What can this exercise tell you in terms of rough “guesstimates” versus carefully calculated approximations?

Exercise:

Check Your Understanding

Problem:

Using mental math and your understanding of fundamental units, approximate the area of a regulation basketball court. Describe the process you used to arrive at your final approximation.

Solution:

An average male is about two meters tall. It would take approximately 15 men laid out end to end to cover the length, and about 7 to cover the width. That gives an approximate area of $420\ \text{m}^2$.

Summary

Scientists often approximate the values of quantities to perform calculations and analyze systems.

Problems & Exercises

Exercise:

Problem: How many heartbeats are there in a lifetime?

Solution:

Sample answer: 2×10^9 heartbeats

Exercise:

Problem:

A generation is about one-third of a lifetime. Approximately how many generations have passed since the year 0 AD?

Exercise:

Problem:

How many times longer than the mean life of an extremely unstable atomic nucleus is the lifetime of a human? (Hint: The lifetime of an unstable atomic nucleus is on the order of 10^{-22} s.)

Solution:

Sample answer: 2×10^{31} if an average human lifetime is taken to be about 70 years.

Exercise:

Problem:

Calculate the approximate number of atoms in a bacterium. Assume that the average mass of an atom in the bacterium is ten times the mass of a hydrogen atom. (Hint: The mass of a hydrogen atom is on the order of 10^{-27} kg and the mass of a bacterium is on the order of 10^{-15} kg.)

This color-enhanced photo shows *Salmonella typhimurium* (red) attacking human cells. These bacteria are commonly known for causing foodborne illness. Can you estimate the number of atoms in each bacterium? (credit: Rocky Mountain Laboratories, NIAID, NIH)

Exercise:

Problem:

Approximately how many atoms thick is a cell membrane, assuming all atoms there average about twice the size of a hydrogen atom?

Solution:

Sample answer: 50 atoms

Exercise:

Problem:

(a) What fraction of Earth’s diameter is the greatest ocean depth? (b) The greatest mountain height?

Exercise:

Problem:

(a) Calculate the number of cells in a hummingbird assuming the mass of an average cell is ten times the mass of a bacterium. (b) Making the same assumption, how many cells are there in a human?

Solution:

Sample answers:

(a) 10^{12} cells/hummingbird

(b) 10^{16} cells/human

Exercise:

Problem:

Assuming one nerve impulse must end before another can begin, what is the maximum firing rate of a nerve in impulses per second?

Glossary

approximation

an estimated value based on prior experience and reasoning

Introduction to Work, Energy, and Energy Resources class="introduction"

How many forms of energy can you identify in this photograph of a wind farm in Iowa? (credit: Jürgen from Sandesneben, Germany, Wikimedia Commons)

Energy plays an essential role both in everyday events and in scientific phenomena. You can no doubt name many forms of energy, from that provided by our foods, to the energy we use to run our cars, to the sunlight that warms us on the beach. You can also cite examples of what people call energy that may not be scientific, such as someone having an energetic personality. Not only does energy have many interesting forms, it is involved in almost all phenomena, and is one of the most important concepts of physics. What makes it even more important is that the total amount of energy in the universe is constant. Energy can change forms, but it cannot appear from nothing or disappear without a trace. Energy is thus one of a handful of physical quantities that we say is *conserved*.

Conservation of energy (as physicists like to call the principle that energy can neither be created nor destroyed) is based on experiment. Even as scientists discovered new forms of energy, conservation of energy has always been found to apply. Perhaps the most dramatic example of this was supplied by Einstein when he suggested that mass is equivalent to energy (his famous equation $E = \text{mc}^2$).

From a societal viewpoint, energy is one of the major building blocks of modern civilization. Energy resources are key limiting factors to economic growth. The world use of energy resources, especially oil, continues to grow, with ominous consequences economically, socially, politically, and environmentally. We will briefly examine the world's energy use patterns at the end of this chapter.

There is no simple, yet accurate, scientific definition for energy. Energy is characterized by its many forms and the fact that it is conserved. We can loosely define **energy** as the ability to do work, admitting that in some circumstances not all energy is available to do work. Because of the association of energy with work, we begin the chapter with a discussion of work. Work is intimately related to energy and how energy moves from one system to another or changes form.

Work: The Scientific Definition

- Explain how an object must be displaced for a force on it to do work.
- Explain how relative directions of force and displacement determine whether the work done is positive, negative, or zero.

What It Means to Do Work

The scientific definition of work differs in some ways from its everyday meaning. Certain things we think of as hard work, such as writing an exam or carrying a heavy load on level ground, are not work as defined by a scientist. The scientific definition of work reveals its relationship to energy—whenever work is done, energy is transferred.

For work, in the scientific sense, to be done, a force must be exerted and there must be displacement in the direction of the force.

Formally, the **work** done on a system by a constant force is defined to be *the product of the component of the force in the direction of motion times the distance through which the force acts*. For one-way motion in one dimension, this is expressed in equation form as
Equation:

$$W = \mid \mathbf{F} \mid (\cos \theta) \mid \mathbf{d} \mid,$$

where W is work, $\mathbf{d}$ is the displacement of the system, and θ is the angle between the force vector $\mathbf{F}$ and the displacement vector $\mathbf{d}$, as in [link]. We can also write this as
Equation:

$$W = \text{Fd} \cos \theta.$$

To find the work done on a system that undergoes motion that is not one-way or that is in two or three dimensions, we divide the motion into one-way one-dimensional segments and add up the work done over each segment.

Note:

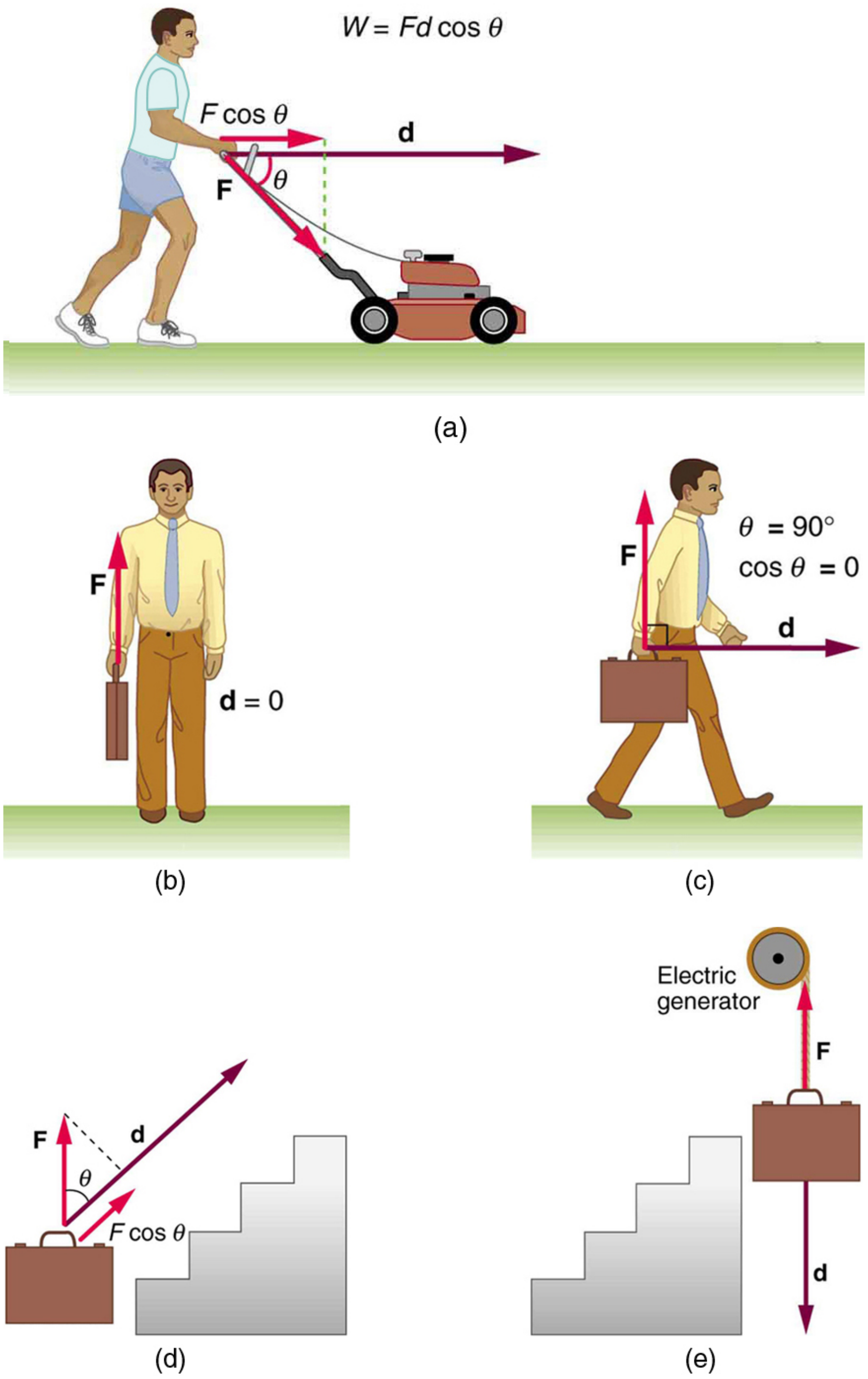
What is Work?

The work done on a system by a constant force is *the product of the component of the force in the direction of motion times the distance through which the force acts*. For one-way motion in one dimension, this is expressed in equation form as

Equation:

$$W = \mathrm{Fd}\cos\theta,$$

where W is work, F is the magnitude of the force on the system, d is the magnitude of the displacement of the system, and θ is the angle between the force vector $\mathbf{F}$ and the displacement vector $\mathbf{d}$.



Examples of work. (a) The work done by the force $\mathbf{F}$ on this lawn mower is $\mathrm{Fd}\cos\theta$. Note that $F\cos\theta$ is the component of the force in the direction of motion. (b) A person holding a briefcase does no work on it, because there is no displacement. No energy is transferred to or from the briefcase. (c) The person moving the briefcase horizontally at a constant speed does no work on it, and transfers no energy to it. (d) Work *is* done on the briefcase by carrying it up stairs at constant speed, because there is necessarily a component of force $\mathbf{F}$ in the direction of the motion. Energy is transferred to the briefcase and could in turn be used to do work. (e) When the briefcase is lowered, energy is transferred out of the briefcase and into an electric generator. Here the work done on the briefcase by the generator is negative, removing energy from the briefcase, because $\mathbf{F}$ and $\mathbf{d}$ are in opposite directions.

To examine what the definition of work means, let us consider the other situations shown in [link]. The person holding the briefcase in [link](b) does no work, for example. Here $d = 0$, so $W = 0$. Why is it you get tired just holding a load? The answer is that your muscles are doing work against one another, *but they are doing no work on the system of interest* (the “briefcase-Earth system”—see Gravitational Potential Energy for more details). There must be displacement for work to be done, and there must be a component of the force in the direction of the motion. For example, the person carrying the briefcase on level ground in [link](c) does no work on it, because the force is perpendicular to the motion. That is, $\cos 90^{\text{o}} = 0$, and so $W = 0$.

In contrast, when a force exerted on the system has a component in the direction of motion, such as in [link](d), work *is* done—energy is transferred to the briefcase. Finally, in [link](e), energy is transferred from the briefcase to a generator. There are two good ways to interpret this energy transfer. One interpretation is that the briefcase’s weight does work on the generator, giving it energy. The other interpretation is that the generator does negative work on the briefcase, thus removing energy from it. The drawing shows the latter, with the force from the generator upward on the briefcase, and the displacement downward. This makes $\theta = 180^\circ$, and $\cos 180^\circ = -1$; therefore, W is negative.

Calculating Work

Work and energy have the same units. From the definition of work, we see that those units are force times distance. Thus, in SI units, work and energy are measured in **newton-meters**. A newton-meter is given the special name **joule** (J), and $1\text{ J} = 1\text{ N}\cdot\text{m} = 1\text{ kg}\cdot\text{m}^2/\text{s}^2$. One joule is not a large amount of energy; it would lift a small 100-gram apple a distance of about 1 meter.

Example:
Calculating the Work You Do to Push a Lawn Mower Across a Large Lawn

How much work is done on the lawn mower by the person in [link](a) if he exerts a constant force of 75.0 N at an angle 35° below the horizontal and pushes the mower 25.0 m on level ground? Convert the amount of work from joules to kilocalories and compare it with this person's average daily intake of $10,000\text{ kJ}$ (about 2400 kcal) of food energy. One *calorie* (1 cal) of heat is the amount required to warm 1 g of water by 1°C, and is equivalent to 4.184 J, while one *food calorie* (1 kcal) is equivalent to 4184 J.

Strategy

We can solve this problem by substituting the given values into the definition of work done on a system, stated in the equation $W = \text{Fd}\cos\theta$. The force, angle, and displacement are given, so that only the work W is unknown.

Solution

The equation for the work is

Equation:

$$W = \text{Fd}\cos\theta.$$

Substituting the known values gives

Equation:

$$\begin{aligned} W &= (75.0\text{ N})(25.0\text{ m})\cos\left(35.0^{\circ}\right) \\ &= 1536\text{ J} = 1.54 \times 10^{3}\text{ J}. \end{aligned}$$

Converting the work in joules to kilocalories yields $W = (1536\text{ J})(1\text{ kcal}/4184\text{ J}) = 0.367\text{ kcal}$. The ratio of the work done to the daily consumption is

Equation:

$$\frac{W}{2400\text{ kcal}} = 1.53 \times 10^{-4}.$$

Discussion

This ratio is a tiny fraction of what the person consumes, but it is typical. Very little of the energy released in the consumption of food is used to do work. Even when we “work” all day long, less than 10% of our food energy intake is used to do work and more than 90% is converted to thermal energy or stored as chemical energy in fat.

Section Summary

- Work is the transfer of energy by a force acting on an object as it is displaced.
- The work W that a force $\mathbf{F}$ does on an object is the product of the magnitude F of the force, times the magnitude d of the displacement, times the cosine of the angle θ between them. In symbols,

  Equation:

  $$W = \text{Fd}\cos\theta.$$

- The SI unit for work and energy is the joule (J), where $1\text{ J} = 1\text{ N}\cdot\text{m} = 1\text{ kg}\cdot\text{m}^2/\text{s}^2$.
- The work done by a force is zero if the displacement is either zero or perpendicular to the force.

- The work done is positive if the force and displacement have the same direction, and negative if they have opposite direction.

Conceptual Questions

Exercise:

Problem:

Give an example of something we think of as work in everyday circumstances that is not work in the scientific sense. Is energy transferred or changed in form in your example? If so, explain how this is accomplished without doing work.

Exercise:

Problem:

Give an example of a situation in which there is a force and a displacement, but the force does no work. Explain why it does no work.

Exercise:

Problem:

Describe a situation in which a force is exerted for a long time but does no work. Explain.

Problems & Exercises

Exercise:

Problem:

How much work does a supermarket checkout attendant do on a can of soup he pushes 0.600 m horizontally with a force of 5.00 N? Express your answer in joules and kilocalories.

Solution:

Equation:

$$3.00 \text{ J} = 7.17 \times 10^{-4} \text{ kcal}$$

Exercise:

Problem:

A 75.0-kg person climbs stairs, gaining 2.50 meters in height. Find the work done to accomplish this task.

Exercise:

Problem:

(a) Calculate the work done on a 1500-kg elevator car by its cable to lift it 40.0 m at constant speed, assuming friction averages 100 N. (b) What is the work done on the lift by the gravitational force in this process? (c) What is the total work done on the lift?

Solution:

(a) $5.92 \times 10^5 \text{ J}$

(b) $-5.88 \times 10^5 \text{ J}$

(c) The net force is zero.

Exercise:

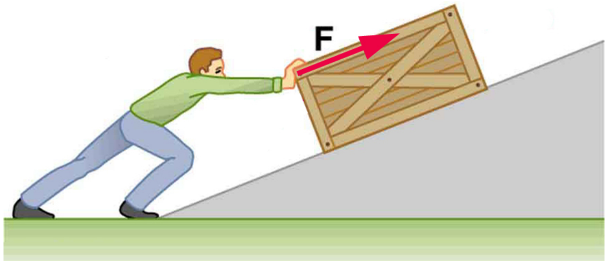
Problem:

Suppose a car travels 108 km at a speed of 30.0 m/s, and uses 2.0 gal of gasoline. Only 30% of the gasoline goes into useful work by the force that keeps the car moving at constant speed despite friction. (See [link] for the energy content of gasoline.) (a) What is the magnitude of the force exerted to keep the car moving at constant speed? (b) If the required force is directly proportional to speed, how many gallons will be used to drive 108 km at a speed of 28.0 m/s?

Exercise:

Problem:

Calculate the work done by an 85.0-kg man who pushes a crate 4.00 m up along a ramp that makes an angle of 20.0° with the horizontal. (See [link].) He exerts a force of 500 N on the crate parallel to the ramp and moves at a constant speed. Be certain to include the work he does on the crate *and* on his body to get up the ramp.



A man pushes a crate up a ramp.

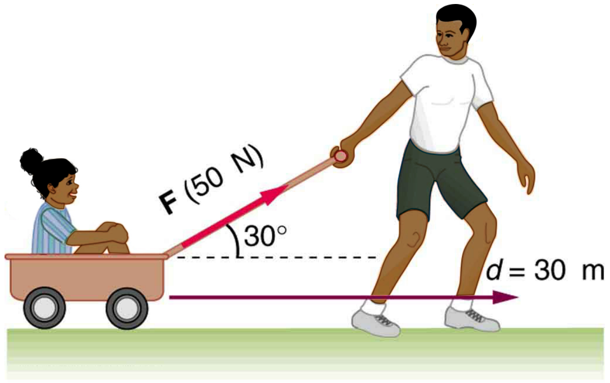
Solution:
Equation:

$$3.14 \times 10^3 \text{ J}$$

Exercise:

Problem:

How much work is done by the boy pulling his sister 30.0 m in a wagon as shown in [link]? Assume no friction acts on the wagon.



The boy does work on the system of the wagon and the child when he pulls them as shown.

Exercise:

Problem:

A shopper pushes a grocery cart 20.0 m at constant speed on level ground, against a 35.0 N frictional force. He pushes in a direction 25.0º below the horizontal. (a) What is the work done on the cart by friction? (b) What is the work done on the cart by the gravitational force? (c) What is the work done on the cart by the shopper? (d) Find the force the shopper exerts, using energy considerations. (e) What is the total work done on the cart?

Solution:

(a) -700 J

(b) 0

(c) 700 J

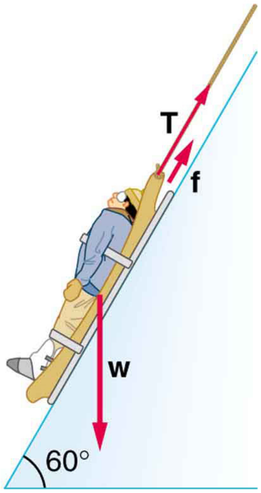
(d) 38.6 N

(e) 0

Exercise:

Problem:

Suppose the ski patrol lowers a rescue sled and victim, having a total mass of 90.0 kg, down a 60.0º slope at constant speed, as shown in [link]. The coefficient of friction between the sled and the snow is 0.100. (a) How much work is done by friction as the sled moves 30.0 m along the hill? (b) How much work is done by the rope on the sled in this distance? (c) What is the work done by the gravitational force on the sled? (d) What is the total work done?



A rescue sled and victim are lowered down a steep slope.

Glossary

energy

the ability to do work

work

the transfer of energy by a force that causes an object to be displaced; the product of the component of the force in the direction of the displacement and the magnitude of the displacement

joule

SI unit of work and energy, equal to one newton-meter

Kinetic Energy and the Work-Energy Theorem

- Explain work as a transfer of energy and net work as the work done by the net force.
- Explain and apply the work-energy theorem.

Work Transfers Energy

What happens to the work done on a system? Energy is transferred into the system, but in what form? Does it remain in the system or move on? The answers depend on the situation. For example, if the lawn mower in [link] (a) is pushed just hard enough to keep it going at a constant speed, then energy put into the mower by the person is removed continuously by friction, and eventually leaves the system in the form of heat transfer. In contrast, work done on the briefcase by the person carrying it up stairs in [link](d) is stored in the briefcase-Earth system and can be recovered at any time, as shown in [link](e). In fact, the building of the pyramids in ancient Egypt is an example of storing energy in a system by doing work on the system. Some of the energy imparted to the stone blocks in lifting them during construction of the pyramids remains in the stone-Earth system and has the potential to do work.

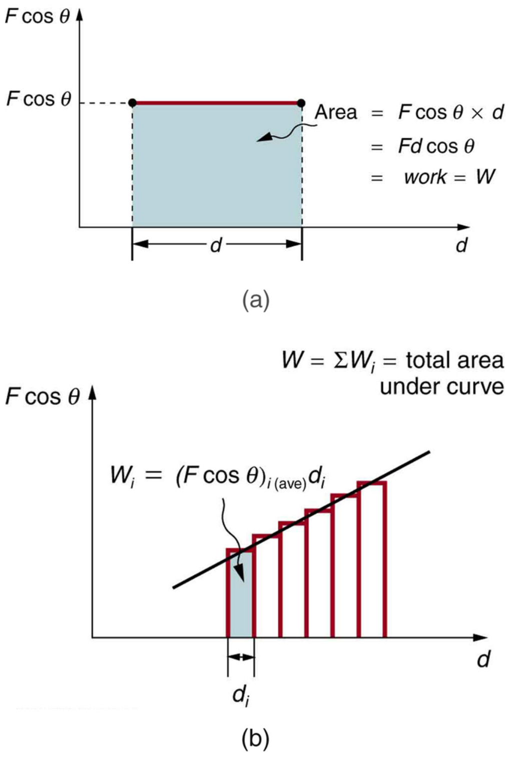
In this section we begin the study of various types of work and forms of energy. We will find that some types of work leave the energy of a system constant, for example, whereas others change the system in some way, such as making it move. We will also develop definitions of important forms of energy, such as the energy of motion.

Net Work and the Work-Energy Theorem

We know from the study of Newton’s laws in Dynamics: Force and Newton's Laws of Motion that net force causes acceleration. We will see in this section that work done by the net force gives a system energy of motion, and in the process we will also find an expression for the energy of motion.

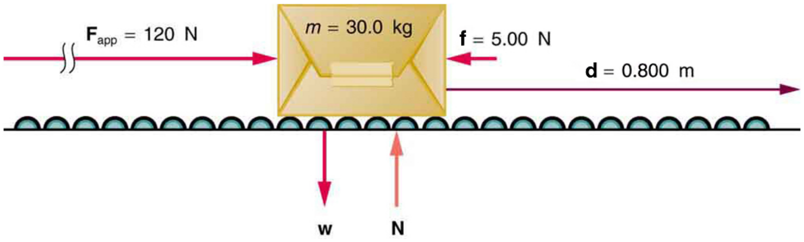
Let us start by considering the total, or net, work done on a system. Net work is defined to be the sum of work done by all external forces—that is, **net work** is the work done by the net external force $\mathbf{F}_{\text{net}}$. In equation form, this is $W_{\text{net}} = F_{\text{net}} d \cos \theta$ where θ is the angle between the force vector and the displacement vector.

[link](a) shows a graph of force versus displacement for the component of the force in the direction of the displacement—that is, an $F \cos \theta$ vs. d graph. In this case, $F \cos \theta$ is constant. You can see that the area under the graph is $Fd \cos \theta$, or the work done. [link](b) shows a more general process where the force varies. The area under the curve is divided into strips, each having an average force $(F \cos \theta)_{i(\text{ave})}$. The work done is $(F \cos \theta)_{i(\text{ave})} d_i$ for each strip, and the total work done is the sum of the W_i. Thus the total work done is the total area under the curve, a useful property to which we shall refer later.



(a) A graph of $F \cos \theta$ vs. d, when $F \cos \theta$ is constant. The area under the curve represents the work done by the force. (b) A graph of F cos θ vs. d in which the force varies. The work done for each interval is the area of each strip; thus, the total area under the curve equals the total work done.

Net work will be simpler to examine if we consider a one-dimensional situation where a force is used to accelerate an object in a direction parallel to its initial velocity. Such a situation occurs for the package on the roller belt conveyor system shown in [link].



A package on a roller belt is pushed horizontally through a distance $\mathbf{d}$.

The force of gravity and the normal force acting on the package are perpendicular to the displacement and do no work. Moreover, they are also equal in magnitude and opposite in direction so they cancel in calculating the net force. The net force arises solely from the horizontal applied force $\mathbf{F}_{\text{app}}$ and the horizontal friction force $\mathbf{f}$. Thus, as expected, the net force is parallel to the displacement, so that $\theta = 0^{\mathrm{o}}$ and $\cos\theta = 1$, and the net work is given by

Equation:

$$W_{\text{net}} = F_{\text{net}}d.$$

The effect of the net force $\mathbf{F}_{\text{net}}$ is to accelerate the package from v_0 to v. The kinetic energy of the package increases, indicating that the net work done on the system is positive. (See [link].) By using Newton’s second law, and doing some algebra, we can reach an interesting conclusion. Substituting $F_{\text{net}} = \text{ma}$ from Newton’s second law gives

Equation:

$$W_{\text{net}} = \text{mad}.$$

To get a relationship between net work and the speed given to a system by the net force acting on it, we take $d = x - x_0$ and use the equation studied in Motion Equations for Constant Acceleration in One Dimension for the change in speed over a distance d if the acceleration has the constant value a; namely, $v^2 = {v_0}^2 + 2\text{ad}$ (note that a appears in the expression for the net work). Solving for acceleration gives $a = \frac{v^2 - {v_0}^2}{2d}$. When a is substituted into the preceding expression for W_{net}, we obtain

Equation:

$$W_{\text{net}} = m\left(\frac{v^2 - {v_0}^2}{2d}\right)d.$$

The d cancels, and we rearrange this to obtain

Equation:

$$W_{\text{ net}} = \frac{1}{2}mv^2 - \frac{1}{2}mv_0^2.$$

This expression is called the **work-energy theorem**, and it actually applies *in general* (even for forces that vary in direction and magnitude), although we have derived it for the special case of a constant force parallel to the displacement. The theorem implies that the net work on a system equals the change in the quantity $\frac{1}{2}mv^2$. This quantity is our first example of a form of energy.

Note:
The Work-Energy Theorem
The net work on a system equals the change in the quantity $\frac{1}{2}mv^2$.
Equation:

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}\text{mv}_0^2$$

The quantity $\frac{1}{2}mv^2$ in the work-energy theorem is defined to be the translational **kinetic energy** (KE) of a mass m moving at a speed v. (*Translational* kinetic energy is distinct from *rotational* kinetic energy, which is considered later.) In equation form, the translational kinetic energy,
Equation:

$$\text{KE} = \frac{1}{2}mv^2,$$

is the energy associated with translational motion. Kinetic energy is a form of energy associated with the motion of a particle, single body, or system of objects moving together.

We are aware that it takes energy to get an object, like a car or the package in [link], up to speed, but it may be a bit surprising that kinetic energy is proportional to speed squared. This proportionality means, for example, that a car traveling at 100 km/h has four times the kinetic energy it has at 50 km/h, helping to explain why high-speed collisions are so devastating. We will now consider a series of examples to illustrate various aspects of work and energy.

Example:

Calculating the Kinetic Energy of a Package

Suppose a 30.0-kg package on the roller belt conveyor system in [link] is moving at 0.500 m/s. What is its kinetic energy?

Strategy

Because the mass m and speed v are given, the kinetic energy can be calculated from its definition as given in the equation $\text{KE}=\frac{1}{2}mv^2$.

Solution

The kinetic energy is given by

Equation:

$$\text{KE}=\frac{1}{2}mv^2.$$

Entering known values gives

Equation:

$$\text{KE}=0.5(30.0\text{ kg})(0.500\text{ m/s})^2,$$

which yields

Equation:

$$\text{KE}=3.75\text{ kg}\cdot\text{m}^2/\text{s}^2=3.75\text{ J}.$$

Discussion

Note that the unit of kinetic energy is the joule, the same as the unit of work, as mentioned when work was first defined. It is also interesting that, although this is a fairly massive package, its kinetic energy is not large at this relatively low speed. This fact is consistent with the observation that people can move packages like this without exhausting themselves.

Example:
Determining the Work to Accelerate a Package

Suppose that you push on the 30.0-kg package in [link] with a constant force of 120 N through a distance of 0.800 m, and that the opposing friction force averages 5.00 N.

(a) Calculate the net work done on the package. (b) Solve the same problem as in part (a), this time by finding the work done by each force that contributes to the net force.

Strategy and Concept for (a)

This is a motion in one dimension problem, because the downward force (from the weight of the package) and the normal force have equal magnitude and opposite direction, so that they cancel in calculating the net force, while the applied force, friction, and the displacement are all horizontal. (See [link].) As expected, the net work is the net force times distance.

Solution for (a)

The net force is the push force minus friction, or $F_{\text{net}} = 120\text{ N} – 5.00\text{ N} = 115\text{ N}$. Thus the net work is

Equation:

$$\begin{aligned} W_{\text{net}} &= F_{\text{net}}d = (115\text{ N})(0.800\text{ m}) \\ &= 92.0\text{ N}\cdot\text{m} = 92.0\text{ J}. \end{aligned}$$

Discussion for (a)

This value is the net work done on the package. The person actually does more work than this, because friction opposes the motion. Friction does negative work and removes some of the energy the person expends and converts it to thermal energy. The net work equals the sum of the work done by each individual force.

Strategy and Concept for (b)

The forces acting on the package are gravity, the normal force, the force of friction, and the applied force. The normal force and force of gravity are each perpendicular to the displacement, and therefore do no work.

Solution for (b)

The applied force does work.

Equation:

$$\begin{aligned} W_{\text{app}} &= F_{\text{app}} d \cos(0^{\text{o}}) = F_{\text{app}} d \\ &= (120 \text{ N})(0.800 \text{ m}) \\ &= 96.0 \text{ J} \end{aligned}$$

The friction force and displacement are in opposite directions, so that $\theta = 180^{\text{o}}$, and the work done by friction is

Equation:

$$\begin{aligned} W_{\text{fr}} &= F_{\text{fr}} d \cos(180^{\text{o}}) = -F_{\text{fr}} d \\ &= -(5.00 \text{ N})(0.800 \text{ m}) \\ &= -4.00 \text{ J}. \end{aligned}$$

So the amounts of work done by gravity, by the normal force, by the applied force, and by friction are, respectively,

Equation:

$$\begin{aligned} W_{\text{gr}} &= 0, \\ W_{\text{N}} &= 0, \\ W_{\text{app}} &= 96.0 \text{ J}, \\ W_{\text{fr}} &= -4.00 \text{ J}. \end{aligned}$$

The total work done as the sum of the work done by each force is then seen to be

Equation:

$$W_{\text{total}} = W_{\text{gr}} + W_{\text{N}} + W_{\text{app}} + W_{\text{fr}} = 92.0 \text{ J}.$$

Discussion for (b)

The calculated total work W_{total} as the sum of the work by each force agrees, as expected, with the work W_{net} done by the net force. The work done by a collection of forces acting on an object can be calculated by either approach.

Example:

Determining Speed from Work and Energy

Find the speed of the package in [link] at the end of the push, using work and energy concepts.

Strategy

Here the work-energy theorem can be used, because we have just calculated the net work, W_{net}, and the initial kinetic energy, $\frac{1}{2}m{v_0}^2$. These calculations allow us to find the final kinetic energy, $\frac{1}{2}mv^2$, and thus the final speed v.

Solution

The work-energy theorem in equation form is

Equation:

$$W_{\text{net}}=\frac{1}{2}mv^2-\frac{1}{2}m{v_0}^2.$$

Solving for $\frac{1}{2}mv^2$ gives

Equation:

$$\frac{1}{2}\text{mv}^2=W_{\text{net}}+\frac{1}{2}m{v_0}^2.$$

Thus,

Equation:

$$\frac{1}{2}mv^2=92.0\text{ J}+3.75\text{ J}=95.75\text{ J}.$$

Solving for the final speed as requested and entering known values gives

Equation:

$$\begin{aligned}v &= \sqrt{\frac{2(95.75\text{ J})}{m}}=\sqrt{\frac{191.5\text{ kg}\cdot\text{m}^2/\text{s}^2}{30.0\text{ kg}}}\\ &= 2.53\text{ m/s}.\end{aligned}$$

Discussion

Using work and energy, we not only arrive at an answer, we see that the final kinetic energy is the sum of the initial kinetic energy and the net work done on the package. This means that the work indeed adds to the energy of the package.

Example:

Work and Energy Can Reveal Distance, Too

How far does the package in [link] coast after the push, assuming friction remains constant? Use work and energy considerations.

Strategy

We know that once the person stops pushing, friction will bring the package to rest. In terms of energy, friction does negative work until it has removed all of the package’s kinetic energy. The work done by friction is the force of friction times the distance traveled times the cosine of the angle between the friction force and displacement; hence, this gives us a way of finding the distance traveled after the person stops pushing.

Solution

The normal force and force of gravity cancel in calculating the net force. The horizontal friction force is then the net force, and it acts opposite to the displacement, so $\theta = 180^{\text{o}}$. To reduce the kinetic energy of the package to zero, the work W_{fr} by friction must be minus the kinetic energy that the package started with plus what the package accumulated due to the pushing. Thus $W_{\text{fr}} = -95.75\text{ J}$. Furthermore, $W_{\text{fr}} = fd\prime \cos \theta = -fd\prime$, where $d\prime$ is the distance it takes to stop. Thus,

Equation:

$$d\prime = -\frac{W_{\text{fr}}}{f} = -\frac{-95.75\text{ J}}{5.00\text{ N}},$$

and so

Equation:

$$d\prime = 19.2\text{ m}.$$

Discussion

This is a reasonable distance for a package to coast on a relatively friction-free conveyor system. Note that the work done by friction is negative (the force is in the opposite direction of motion), so it removes the kinetic energy.

Some of the examples in this section can be solved without considering energy, but at the expense of missing out on gaining insights about what work and energy are doing in this situation. On the whole, solutions involving energy are generally shorter and easier than those using kinematics and dynamics alone.

Section Summary

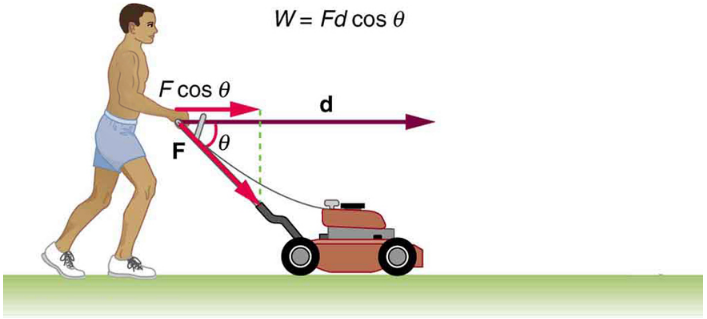
- The net work W_{net} is the work done by the net force acting on an object.
- Work done on an object transfers energy to the object.
- The translational kinetic energy of an object of mass m moving at speed v is $\text{KE} = \frac{1}{2}mv^2$.
- The work-energy theorem states that the net work W_{net} on a system changes its kinetic energy, $W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}m{v_0}^2$.

Conceptual Questions

Exercise:

Problem:

The person in [link] does work on the lawn mower. Under what conditions would the mower gain energy? Under what conditions would it lose energy?



Exercise:

Problem:

Work done on a system puts energy into it. Work done by a system removes energy from it. Give an example for each statement.

Exercise:

Problem:

When solving for speed in [link], we kept only the positive root. Why?

Problems & Exercises

Exercise:

Problem:

Compare the kinetic energy of a 20,000-kg truck moving at 110 km/h with that of an 80.0-kg astronaut in orbit moving at 27,500 km/h.

Solution:

$1/250$

Exercise:

Problem:

(a) How fast must a 3000-kg elephant move to have the same kinetic energy as a 65.0-kg sprinter running at 10.0 m/s? (b) Discuss how the larger energies needed for the movement of larger animals would relate to metabolic rates.

Exercise:

Problem:

Confirm the value given for the kinetic energy of an aircraft carrier in [link]. You will need to look up the definition of a nautical mile (1 knot = 1 nautical mile/h).

Solution:

1.1×10^{10} J

Exercise:

Problem:

(a) Calculate the force needed to bring a 950-kg car to rest from a speed of 90.0 km/h in a distance of 120 m (a fairly typical distance for a non-panic stop). (b) Suppose instead the car hits a concrete abutment at full speed and is brought to a stop in 2.00 m. Calculate the force exerted on the car and compare it with the force found in part (a).

Exercise:

Problem:

A car’s bumper is designed to withstand a 4.0-km/h (1.1-m/s) collision with an immovable object without damage to the body of the car. The bumper cushions the shock by absorbing the force over a distance. Calculate the magnitude of the average force on a bumper that collapses 0.200 m while bringing a 900-kg car to rest from an initial speed of 1.1 m/s.

Solution:

2.8×10^{3} N

Exercise:

Problem:

Boxing gloves are padded to lessen the force of a blow. (a) Calculate the force exerted by a boxing glove on an opponent's face, if the glove and face compress 7.50 cm during a blow in which the 7.00-kg arm and glove are brought to rest from an initial speed of 10.0 m/s. (b) Calculate the force exerted by an identical blow in the gory old days when no gloves were used and the knuckles and face would compress only 2.00 cm. (c) Discuss the magnitude of the force with glove on. Does it seem high enough to cause damage even though it is lower than the force with no glove?

Exercise:

Problem:

Using energy considerations, calculate the average force a 60.0-kg sprinter exerts backward on the track to accelerate from 2.00 to 8.00 m/s in a distance of 25.0 m, if he encounters a headwind that exerts an average force of 30.0 N against him.

Solution:

102 N

Glossary

net work
: work done by the net force, or vector sum of all the forces, acting on an object

work-energy theorem
: the result, based on Newton's laws, that the net work done on an object is equal to its change in kinetic energy

kinetic energy

the energy an object has by reason of its motion, equal to $\frac{1}{2}mv^2$ for the translational (i.e., non-rotational) motion of an object of mass m moving at speed v

Gravitational Potential Energy

- Explain gravitational potential energy in terms of work done against gravity.
- Show that the gravitational potential energy of an object of mass m at height h on Earth is given by $\mathrm{PE_g} = \mathrm{mgh}$.
- Show how knowledge of the potential energy as a function of position can be used to simplify calculations and explain physical phenomena.

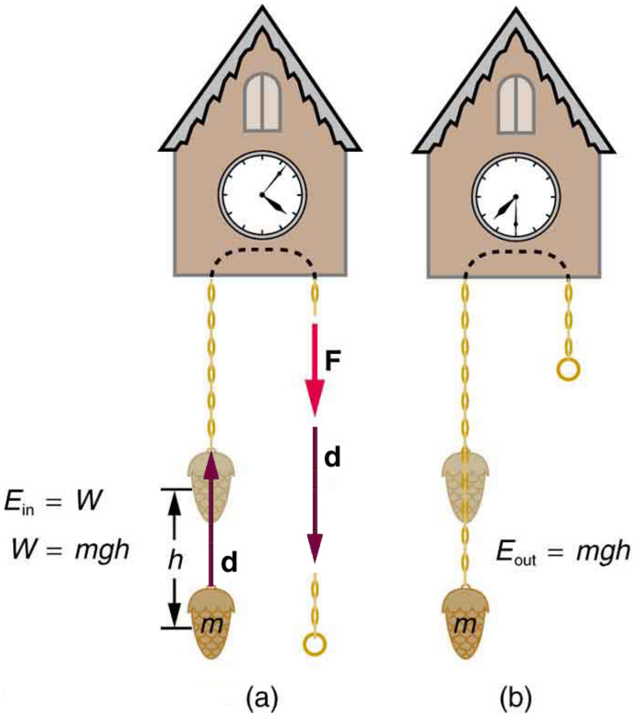
Work Done Against Gravity

Climbing stairs and lifting objects is work in both the scientific and everyday sense—it is work done against the gravitational force. When there is work, there is a transformation of energy. The work done against the gravitational force goes into an important form of stored energy that we will explore in this section.

Let us calculate the work done in lifting an object of mass m through a height h, such as in [link]. If the object is lifted straight up at constant speed, then the force needed to lift it is equal to its weight mg. The work done on the mass is then $\mathrm{W} = \mathrm{Fd} = \mathrm{mgh}$. We define this to be the **gravitational potential energy** $(\mathrm{PE_g})$ put into (or gained by) the object-Earth system. This energy is associated with the state of separation between two objects that attract each other by the gravitational force. For convenience, we refer to this as the $\mathrm{PE_g}$ gained by the object, recognizing that this is energy stored in the gravitational field of Earth. Why do we use the word “system”? Potential energy is a property of a system rather than of a single object—due to its physical position. An object’s gravitational potential is due to its position relative to the surroundings within the Earth-object system. The force applied to the object is an external force, from outside the system. When it does positive work it increases the gravitational potential energy of the system. Because gravitational potential energy depends on relative position, we need a reference level at which to set the potential energy equal to 0. We usually choose this point to be Earth’s surface, but this point is arbitrary; what is important is the *difference* in gravitational potential energy, because this difference is what relates to the work done. The difference in gravitational potential energy of an object (in the Earth-object system) between two rungs of a ladder will be the same for the first two rungs as for the last two rungs.

Converting Between Potential Energy and Kinetic Energy

Gravitational potential energy may be converted to other forms of energy, such as kinetic energy. If we release the mass, gravitational force will do an amount of work equal to mgh on it, thereby increasing its kinetic energy by that same amount (by the work-energy theorem). We will find it more useful to consider just the conversion of PE_{g} to KE without explicitly considering the intermediate step of work. (See [link].) This shortcut makes it is easier to solve problems using energy (if possible) rather than explicitly using forces.



(a) The work done to lift the weight is stored in the mass-Earth system as gravitational potential energy. (b) As the weight moves downward, this gravitational potential energy is transferred to the cuckoo clock.

More precisely, we define the *change* in gravitational potential energy $\Delta\text{PE}_{\text{g}}$ to be

Equation:

$$\Delta\text{PE}_{\text{g}} = \text{mgh},$$

where, for simplicity, we denote the change in height by h rather than the usual Δh. Note that h is positive when the final height is greater than the initial height, and vice versa. For example, if a 0.500-kg mass hung from a cuckoo clock is raised 1.00 m, then its change in gravitational potential energy is

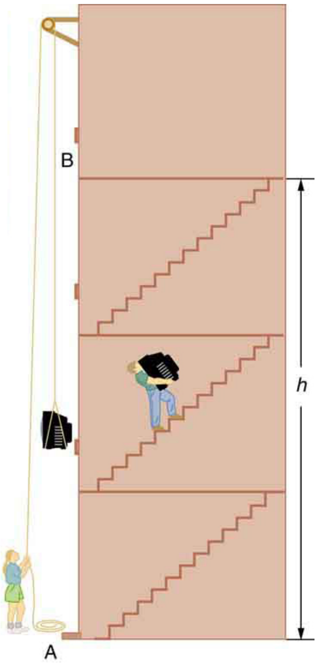
Equation:

$$\begin{array}{lll} mgh & = & (0.500\text{ kg})\left(9.80\text{ m/s}^2\right)(1.00\text{ m}) \\ & = & 4.90\text{ kg}\cdot\text{m}^2/\text{s}^2 = 4.90\text{ J}. \end{array}$$

Note that the units of gravitational potential energy turn out to be joules, the same as for work and other forms of energy. As the clock runs, the mass is lowered. We can think of the mass as gradually giving up its 4.90 J of gravitational potential energy, *without directly considering the force of gravity that does the work*.

Using Potential Energy to Simplify Calculations

The equation $\Delta\text{PE}_\text{g} = \text{mgh}$ applies for any path that has a change in height of h, not just when the mass is lifted straight up. (See [link].) It is much easier to calculate mgh (a simple multiplication) than it is to calculate the work done along a complicated path. The idea of gravitational potential energy has the double advantage that it is very broadly applicable and it makes calculations easier. From now on, we will consider that any change in vertical position h of a mass m is accompanied by a change in gravitational potential energy mgh, and we will avoid the equivalent but more difficult task of calculating work done by or against the gravitational force.



The change in gravitational potential energy $(\Delta \mathrm{PE_g})$ between points A and B is independent of the path. $\Delta \mathrm{PE_g} = \mathrm{mgh}$ for any path between the two points. Gravity is one of a small class of forces where the work done by or against the force depends only on the starting and ending points, not on the path between them.

Example:

The Force to Stop Falling

A 60.0-kg person jumps onto the floor from a height of 3.00 m. If he lands stiffly (with his knee joints compressing by 0.500 cm), calculate the force on the knee joints.

Strategy

This person's energy is brought to zero in this situation by the work done on him by the floor as he stops. The initial PE_{g} is transformed into KE as he falls. The work done by the floor reduces this kinetic energy to zero.

Solution

The work done on the person by the floor as he stops is given by

Equation:

$$W = \text{Fd} \cos \theta = -\text{Fd},$$

with a minus sign because the displacement while stopping and the force from floor are in opposite directions $(\cos \theta = \cos 180^{\text{o}} = -1)$. The floor removes energy from the system, so it does negative work.

The kinetic energy the person has upon reaching the floor is the amount of potential energy lost by falling through height h:

Equation:

$$\text{KE} = -\Delta\text{PE}_{\text{g}} = -\text{mgh},$$

The distance d that the person's knees bend is much smaller than the height h of the fall, so the additional change in gravitational potential energy during the knee bend is ignored.

The work W done by the floor on the person stops the person and brings the person's kinetic energy to zero:

Equation:

$$W = -\text{KE} = \text{mgh}.$$

Combining this equation with the expression for W gives

Equation:

$$-\text{Fd} = \text{mgh}.$$

Recalling that h is negative because the person fell *down*, the force on the knee joints is given by

Equation:

$$F=-\frac{\text{mgh}}{d}=-\frac{(60.0\text{ kg})\left(9.80\text{ m/s}^2\right)(-3.00\text{ m})}{5.00\times 10^{-3}\text{ m}}=3.53\times 10^5\text{ N}.$$

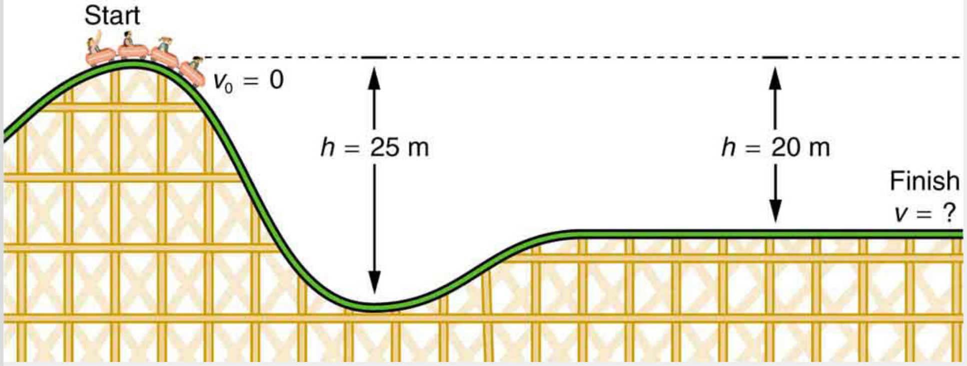
Discussion

Such a large force (500 times more than the person’s weight) over the short impact time is enough to break bones. A much better way to cushion the shock is by bending the legs or rolling on the ground, increasing the time over which the force acts. A bending motion of 0.5 m this way yields a force 100 times smaller than in the example. A kangaroo's hopping shows this method in action. The kangaroo is the only large animal to use hopping for locomotion, but the shock in hopping is cushioned by the bending of its hind legs in each jump.(See [link].)

The work done by the ground upon the kangaroo reduces its kinetic energy to zero as it lands. However, by applying the force of the ground on the hind legs over a longer distance, the impact on the bones is reduced. (credit: Chris Samuel, Flickr)

Example:
Finding the Speed of a Roller Coaster from its Height

(a) What is the final speed of the roller coaster shown in [link] if it starts from rest at the top of the 20.0 m hill and work done by frictional forces is negligible? (b) What is its final speed (again assuming negligible friction) if its initial speed is 5.00 m/s?



The speed of a roller coaster increases as gravity pulls it downhill and is greatest at its lowest point. Viewed in terms of energy, the roller-coaster-Earth system’s gravitational potential energy is converted to kinetic energy. If work done by friction is negligible, all ΔPE_g is converted to KE.

Strategy

The roller coaster loses potential energy as it goes downhill. We neglect friction, so that the remaining force exerted by the track is the normal force, which is perpendicular to the direction of motion and does no work. The net work on the roller coaster is then done by gravity alone. The *loss* of gravitational potential energy from moving *downward* through a distance h equals the *gain* in kinetic energy. This can be written in equation form as $-\Delta\text{PE}_\text{g} = \Delta\text{KE}$. Using the equations for PE_g and KE, we can solve for the final speed v, which is the desired quantity.

Solution for (a)

Here the initial kinetic energy is zero, so that $\Delta\text{KE} = \frac{1}{2}mv^2$. The equation for change in potential energy states that $\Delta\text{PE}_\text{g} = mgh$. Since h is negative in this case, we will rewrite this as $\Delta\text{PE}_\text{g} = -mg \mid h \mid$ to show the minus sign clearly. Thus,

Equation:

$$-\Delta \mathrm{PE}_{\mathrm{g}} = \Delta \mathrm{KE}$$

becomes

Equation:

$$\mathrm{mg} \mid h \mid = \frac{1}{2} m v^2.$$

Solving for v, we find that mass cancels and that

Equation:

$$v = \sqrt{2g \mid h \mid}.$$

Substituting known values,

Equation:

$$\begin{aligned} v &= \sqrt{2\left(9.80 \text{ m/s}^2\right)(20.0 \text{ m})} \\ &= 19.8 \text{ m/s}. \end{aligned}$$

Solution for (b)

Again $-\Delta \mathrm{PE}_{\mathrm{g}} = \Delta \mathrm{KE}$. In this case there is initial kinetic energy, so $\Delta \mathrm{KE} = \frac{1}{2} m v^2 - \frac{1}{2} m {v_0}^2$. Thus,

Equation:

$$\mathrm{mg} \mid h \mid = \frac{1}{2} m v^2 - \frac{1}{2} m {v_0}^2.$$

Rearranging gives

Equation:

$$\frac{1}{2} m v^2 = \mathrm{mg} \mid h \mid + \frac{1}{2} m {v_0}^2.$$

This means that the final kinetic energy is the sum of the initial kinetic energy and the gravitational potential energy. Mass again cancels, and

Equation:

$$v = \sqrt{2g \mid h \mid + {v_0}^2}.$$

This equation is very similar to the kinematics equation $v=\sqrt{{v_0}^2+2ad}$, but it is more general—the kinematics equation is valid only for constant acceleration, whereas our equation above is valid for any path regardless of whether the object moves with a constant acceleration. Now, substituting known values gives

Equation:

$$\begin{aligned} v &= \sqrt{2(9.80\text{ m/s}^2)(20.0\text{ m})+(5.00\text{ m/s})^2} \\ &= 20.4\text{ m/s.} \end{aligned}$$

Discussion and Implications

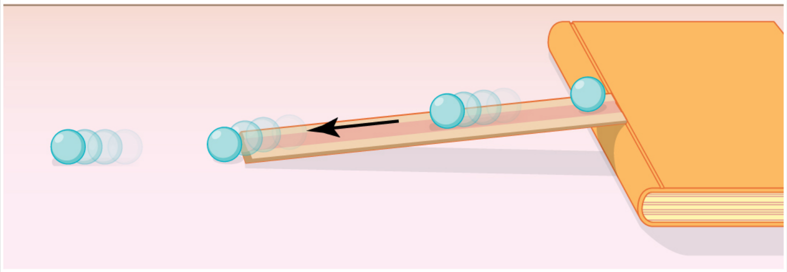
First, note that mass cancels. This is quite consistent with observations made in Falling Objects that all objects fall at the same rate if friction is negligible. Second, only the speed of the roller coaster is considered; there is no information about its direction at any point. This reveals another general truth. When friction is negligible, the speed of a falling body depends only on its initial speed and height, and not on its mass or the path taken. For example, the roller coaster will have the same final speed whether it falls 20.0 m straight down or takes a more complicated path like the one in the figure. Third, and perhaps unexpectedly, the final speed in part (b) is greater than in part (a), but by far less than 5.00 m/s. Finally, note that speed can be found at *any* height along the way by simply using the appropriate value of h at the point of interest.

We have seen that work done by or against the gravitational force depends only on the starting and ending points, and not on the path between, allowing us to define the simplifying concept of gravitational potential energy. We can do the same thing for a few other forces, and we will see that this leads to a formal definition of the law of conservation of energy.

Note:

Making Connections: Take-Home Investigation—Converting Potential to Kinetic Energy

One can study the conversion of gravitational potential energy into kinetic energy in this experiment. On a smooth, level surface, use a ruler of the kind that has a groove running along its length and a book to make an incline (see [link]). Place a marble at the 10-cm position on the ruler and let it roll down the ruler. When it hits the level surface, measure the time it takes to roll one meter. Now place the marble at the 20-cm and the 30-cm positions and again measure the times it takes to roll 1 m on the level surface. Find the velocity of the marble on the level surface for all three positions. Plot velocity squared versus the distance traveled by the marble. What is the shape of each plot? If the shape is a straight line, the plot shows that the marble’s kinetic energy at the bottom is proportional to its potential energy at the release point.



A marble rolls down a ruler, and its speed on the level surface is measured.

Section Summary

- Work done against gravity in lifting an object becomes potential energy of the object-Earth system.
- The change in gravitational potential energy, $\Delta \mathrm{PE_g}$, is $\Delta \mathrm{PE_g} = \mathrm{mgh}$, with h being the increase in height and g the acceleration due to gravity.
- The gravitational potential energy of an object near Earth’s surface is due to its position in the mass-Earth system. Only differences in gravitational potential energy, $\Delta \mathrm{PE_g}$, have physical significance.
- As an object descends without friction, its gravitational potential energy changes into kinetic energy corresponding to increasing speed, so that $\Delta \mathrm{KE} = -\Delta \mathrm{PE_g}$.

Conceptual Questions

Exercise:

Problem:

In [link], we calculated the final speed of a roller coaster that descended 20 m in height and had an initial speed of 5 m/s downhill. Suppose the roller coaster had had an initial speed of 5 m/s *uphill* instead, and it coasted uphill, stopped, and then rolled back down to a final point 20 m below the start. We would find in that case that its final speed is the same as its initial speed. Explain in terms of conservation of energy.

Exercise:

Problem:

Does the work you do on a book when you lift it onto a shelf depend on the path taken? On the time taken? On the height of the shelf? On the mass of the book?

Problems & Exercises

Exercise:

Problem:

A hydroelectric power facility (see [link]) converts the gravitational potential energy of water behind a dam to electric energy. (a) What is the gravitational potential energy relative to the generators of a lake of volume $50.0\ \text{km}^3$ ($\text{mass} = 5.00 \times 10^{13}\ \text{kg}$), given that the lake has an average height of 40.0 m above the generators? (b) Compare this with the energy stored in a 9-megaton fusion bomb.

Hydroelectric facility (credit: Denis

Belevich, Wikimedia Commons)

Solution:

(a) 1.96×10^{16} J

(b) The ratio of gravitational potential energy in the lake to the energy stored in the bomb is 0.52. That is, the energy stored in the lake is approximately half that in a 9-megaton fusion bomb.

Exercise:

Problem:

(a) How much gravitational potential energy (relative to the ground on which it is built) is stored in the Great Pyramid of Cheops, given that its mass is about 7×10^9 kg and its center of mass is 36.5 m above the surrounding ground? (b) How does this energy compare with the daily food intake of a person?

Exercise:

Problem:

Suppose a 350-g kookaburra (a large kingfisher bird) picks up a 75-g snake and raises it 2.5 m from the ground to a branch. (a) How much work did the bird do on the snake? (b) How much work did it do to raise its own center of mass to the branch?

Solution:

(a) 1.8 J

(b) 8.6 J

Exercise:

Problem:

In [link], we found that the speed of a roller coaster that had descended 20.0 m was only slightly greater when it had an initial speed of 5.00 m/s than when it started from rest. This implies that $\Delta\text{PE} >> \text{KE}_\text{i}$. Confirm this statement by taking the ratio of ΔPE to KE_i. (Note that mass cancels.)

Exercise:

Problem:

A 100-g toy car is propelled by a compressed spring that starts it moving. The car follows the curved track in [link]. Show that the final speed of the toy car is 0.687 m/s if its initial speed is 2.00 m/s and it coasts up the frictionless slope, gaining 0.180 m in altitude.

A toy car moves up a sloped track. (credit: Leszek Leszczynski, Flickr)

Solution:
Equation:

$$v_f = \sqrt{2\text{gh} + {v_0}^2} = \sqrt{2(9.80\ \text{m/s}^2)(-0.180\ \text{m}) + (2.00\ \text{m/s})^2} = 0.687\ \text{m/s}$$

Exercise:

Problem:

In a downhill ski race, surprisingly, little advantage is gained by getting a running start. (This is because the initial kinetic energy is small compared with the gain in gravitational potential energy on even small hills.) To demonstrate this, find the final speed and the time taken for a skier who skies 70.0 m along a 30º slope neglecting friction: (a) Starting from rest. (b) Starting with an initial speed of 2.50 m/s. (c) Does the answer surprise you? Discuss why it is still advantageous to get a running start in very competitive events.

Glossary

gravitational potential energy

the energy an object has due to its position in a gravitational field

Conservative Forces and Potential Energy

- Define conservative force, potential energy, and mechanical energy.
- Explain the potential energy of a spring in terms of its compression when Hooke’s law applies.
- Use the work-energy theorem to show how having only conservative forces implies conservation of mechanical energy.

Potential Energy and Conservative Forces

Work is done by a force, and some forces, such as weight, have special characteristics. A **conservative force** is one, like the gravitational force, for which work done by or against it depends only on the starting and ending points of a motion and not on the path taken. We can define a **potential energy** (PE) for any conservative force, just as we did for the gravitational force. For example, when you wind up a toy, an egg timer, or an old-fashioned watch, you do work against its spring and store energy in it. (We treat these springs as ideal, in that we assume there is no friction and no production of thermal energy.) This stored energy is recoverable as work, and it is useful to think of it as potential energy contained in the spring. Indeed, the reason that the spring has this characteristic is that its force is *conservative*. That is, a conservative force results in stored or potential energy. Gravitational potential energy is one example, as is the energy stored in a spring. We will also see how conservative forces are related to the conservation of energy.

Note:
Potential Energy and Conservative Forces
Potential energy is the energy a system has due to position, shape, or configuration. It is stored energy that is completely recoverable.
A conservative force is one for which work done by or against it depends only on the starting and ending points of a motion and not on the path taken.
We can define a potential energy (PE) for any conservative force. The work done against a conservative force to reach a final configuration

depends on the configuration, not the path followed, and is the potential energy added.

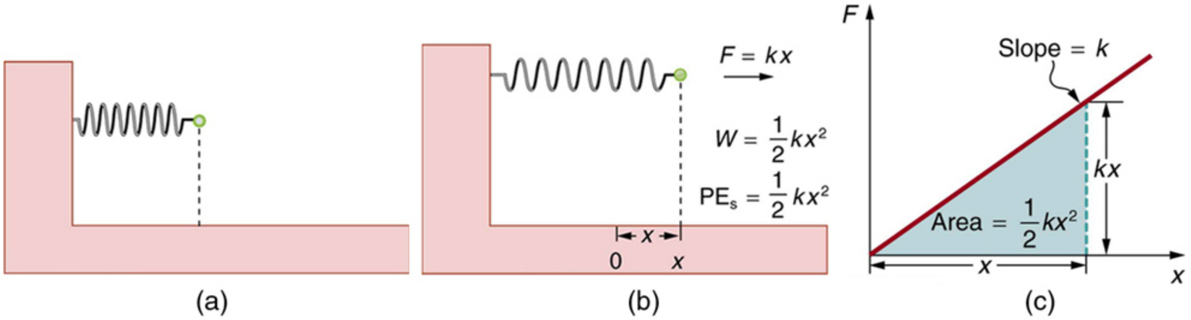
Potential Energy of a Spring

First, let us obtain an expression for the potential energy stored in a spring (PE_{s}). We calculate the work done to stretch or compress a spring that obeys Hooke’s law. (Hooke’s law was examined in Elasticity: Stress and Strain, and states that the magnitude of force F on the spring and the resulting deformation ΔL are proportional, $F = k\Delta L$.) (See [link].) For our spring, we will replace ΔL (the amount of deformation produced by a force F) by the distance x that the spring is stretched or compressed along its length. So the force needed to stretch the spring has magnitude $\mathrm{F} = \mathrm{kx}$, where k is the spring’s force constant. The force increases linearly from 0 at the start to kx in the fully stretched position. The average force is $kx/2$. Thus the work done in stretching or compressing the spring is $W_{\mathrm{s}} = \mathrm{Fd} = \left(\frac{kx}{2}\right)x = \frac{1}{2}kx^2$. Alternatively, we noted in Kinetic Energy and the Work-Energy Theorem that the area under a graph of F vs. x is the work done by the force. In [link](c) we see that this area is also $\frac{1}{2}kx^2$. We therefore define the **potential energy of a spring**, PE_{s}, to be

Equation:

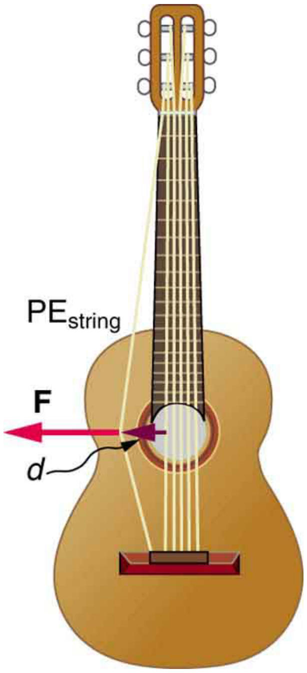
$$\mathrm{PE}_{\mathrm{s}} = \frac{1}{2}\mathrm{kx}^2,$$

where k is the spring’s force constant and x is the displacement from its undeformed position. The potential energy represents the work done *on* the spring and the energy stored in it as a result of stretching or compressing it a distance x. The potential energy of the spring PE_{s} does not depend on the path taken; it depends only on the stretch or squeeze x in the final configuration.



(a) An undeformed spring has no PE_s stored in it. (b) The force needed to stretch (or compress) the spring a distance x has a magnitude $F = kx$, and the work done to stretch (or compress) it is $\frac{1}{2}kx^2$. Because the force is conservative, this work is stored as potential energy (PE_s) in the spring, and it can be fully recovered. (c) A graph of F vs. x has a slope of k, and the area under the graph is $\frac{1}{2}kx^2$. Thus the work done or potential energy stored is $\frac{1}{2}kx^2$.

The equation $\text{PE}_\text{s} = \frac{1}{2}kx^2$ has general validity beyond the special case for which it was derived. Potential energy can be stored in any elastic medium by deforming it. Indeed, the general definition of **potential energy** is energy due to position, shape, or configuration. For shape or position deformations, stored energy is $\text{PE}_\text{s} = \frac{1}{2}kx^2$, where k is the force constant of the particular system and x is its deformation. Another example is seen in [link] for a guitar string.



Work is done to deform the guitar string, giving it potential energy. When released, the potential energy is converted to kinetic energy and back to potential as the string oscillates back and forth. A very small fraction is dissipated as sound energy, slowly removing energy from the string.

Conservation of Mechanical Energy

Let us now consider what form the work-energy theorem takes when only conservative forces are involved. This will lead us to the conservation of energy principle. The work-energy theorem states that the net work done by all forces acting on a system equals its change in kinetic energy. In equation form, this is

Equation:

$$W_{\text{net}} = \frac{1}{2}mv^2 - \frac{1}{2}m{v_0}^2 = \Delta\text{KE}.$$

If only conservative forces act, then

Equation:

$$W_{\text{net}} = W_{\text{c}},$$

where W_{c} is the total work done by all conservative forces. Thus,

Equation:

$$W_{\text{c}} = \Delta\text{KE}.$$

Now, if the conservative force, such as the gravitational force or a spring force, does work, the system loses potential energy. That is, $W_{\text{c}} = -\Delta\text{PE}$. Therefore,

Equation:

$$-\Delta \text{PE} = \Delta \text{KE}$$

or

Equation:

$$\Delta \text{KE} + \Delta \text{PE} = 0.$$

This equation means that the total kinetic and potential energy is constant for any process involving only conservative forces. That is,

Equation:

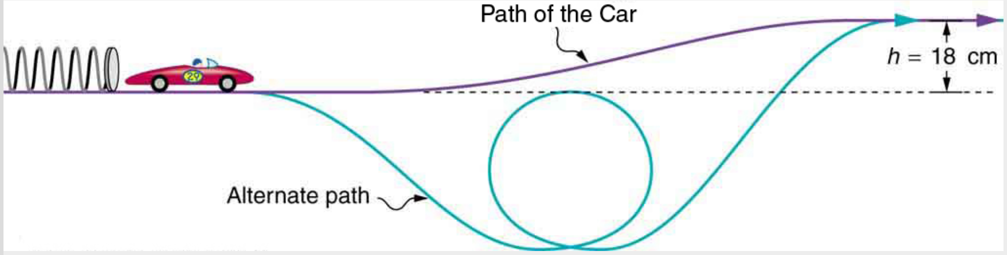
$$\left.\begin{array}{c} \text{KE} + \text{PE} = \text{constant} \\ \text{or} \\ \text{KE}_\text{i} + \text{PE}_\text{i} = \text{KE}_\text{f} + \text{PE}_\text{f} \end{array}\right\} \text{(conservative forces only)},$$

where i and f denote initial and final values. This equation is a form of the work-energy theorem for conservative forces; it is known as the **conservation of mechanical energy** principle. Remember that this applies to the extent that all the forces are conservative, so that friction is negligible. The total kinetic plus potential energy of a system is defined to be its **mechanical energy**, $(\text{KE} + \text{PE})$. In a system that experiences only conservative forces, there is a potential energy associated with each force, and the energy only changes form between KE and the various types of PE, with the total energy remaining constant.

Example:

Using Conservation of Mechanical Energy to Calculate the Speed of a Toy Car

A 0.100-kg toy car is propelled by a compressed spring, as shown in [link]. The car follows a track that rises 0.180 m above the starting point. The spring is compressed 4.00 cm and has a force constant of 250.0 N/m. Assuming work done by friction to be negligible, find (a) how fast the car is going before it starts up the slope and (b) how fast it is going at the top of the slope.



A toy car is pushed by a compressed spring and coasts up a slope. Assuming negligible friction, the potential energy in the spring is first completely converted to kinetic energy, and then to a combination of kinetic and gravitational potential energy as the car rises. The details of the path are unimportant because all forces are conservative—the car would have the same final speed if it took the alternate path shown.

Strategy

The spring force and the gravitational force are conservative forces, so conservation of mechanical energy can be used. Thus,

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} = \mathrm{KE_f} + \mathrm{PE_f}$$

or

Equation:

$$\frac{1}{2}m{v_\mathrm{i}}^2 + mgh_\mathrm{i} + \frac{1}{2}k{x_\mathrm{i}}^2 = \frac{1}{2}m{v_\mathrm{f}}^2 + mgh_\mathrm{f} + \frac{1}{2}k{x_\mathrm{f}}^2,$$

where h is the height (vertical position) and x is the compression of the spring. This general statement looks complex but becomes much simpler when we start considering specific situations. First, we must identify the initial and final conditions in a problem; then, we enter them into the last equation to solve for an unknown.

Solution for (a)

This part of the problem is limited to conditions just before the car is released and just after it leaves the spring. Take the initial height to be zero, so that both h_{i} and h_{f} are zero. Furthermore, the initial speed v_{i} is zero and the final compression of the spring x_{f} is zero, and so several terms in the conservation of mechanical energy equation are zero and it simplifies to

Equation:

$$\frac{1}{2}k{x_{\text{i}}}^{2}=\frac{1}{2}m{v_{\text{f}}}^{2}.$$

In other words, the initial potential energy in the spring is converted completely to kinetic energy in the absence of friction. Solving for the final speed and entering known values yields

Equation:

$$\begin{array}{lll}v_{\text{f}} & = & \sqrt{\frac{k}{m}}x_{\text{i}}\\ & = & \sqrt{\frac{250.0\text{ N/m}}{0.100\text{ kg}}}(0.0400\text{ m})\\ & = & 2.00\text{ m/s}.\end{array}$$

Solution for (b)

One method of finding the speed at the top of the slope is to consider conditions just before the car is released and just after it reaches the top of the slope, completely ignoring everything in between. Doing the same type of analysis to find which terms are zero, the conservation of mechanical energy becomes

Equation:

$$\frac{1}{2}{\text{kx}_{\text{i}}}^{2}=\frac{1}{2}{\text{mv}_{\text{f}}}^{2}+\text{mgh}_{\text{f}}.$$

This form of the equation means that the spring’s initial potential energy is converted partly to gravitational potential energy and partly to kinetic energy. The final speed at the top of the slope will be less than at the bottom. Solving for v_{f} and substituting known values gives

Equation:

$$
\begin{aligned}
v_{\text{f}} &= \sqrt{\frac{k{x_{\text{i}}}^2}{m} - 2gh_{\text{f}}} \\
&= \sqrt{\left(\frac{250.0\ \text{N/m}}{0.100\ \text{kg}}\right)(0.0400\ \text{m})^2 - 2(9.80\ \text{m/s}^2)(0.180\ \text{m})} \\
&= 0.687\ \text{m/s}.
\end{aligned}
$$

Discussion
Another way to solve this problem is to realize that the car’s kinetic energy before it goes up the slope is converted partly to potential energy—that is, to take the final conditions in part (a) to be the initial conditions in part (b).

Note that, for conservative forces, we do not directly calculate the work they do; rather, we consider their effects through their corresponding potential energies, just as we did in [link]. Note also that we do not consider details of the path taken—only the starting and ending points are important (as long as the path is not impossible). This assumption is usually a tremendous simplification, because the path may be complicated and forces may vary along the way.

Note:
PhET Explorations: Energy Skate Park
Learn about conservation of energy with a skater dude! Build tracks, ramps and jumps for the skater and view the kinetic energy, potential energy and friction as he moves. You can also take the skater to different planets or even space!
https://phet.colorado.edu/sims/html/energy-skate-park-basics/latest/energy-skate-park-basics_en.html

Section Summary

- A conservative force is one for which work depends only on the starting and ending points of a motion, not on the path taken.
- We can define potential energy (PE) for any conservative force, just as we defined PE_{g} for the gravitational force.
- The potential energy of a spring is $\mathrm{PE}_{\mathrm{s}} = \frac{1}{2}kx^2$, where k is the spring's force constant and x is the displacement from its undeformed position.
- Mechanical energy is defined to be $\mathrm{KE} + \mathrm{PE}$ for a conservative force.
- When only conservative forces act on and within a system, the total mechanical energy is constant. In equation form,

Equation:

$$\mathrm{KE} + \mathrm{PE} = \text{constant}$$

or

$$\mathrm{KE}_{\mathrm{i}} + \mathrm{PE}_{\mathrm{i}} = \mathrm{KE}_{\mathrm{f}} + \mathrm{PE}_{\mathrm{f}}$$

where i and f denote initial and final values. This is known as the conservation of mechanical energy.

Conceptual Questions

Exercise:

Problem: What is a conservative force?

Exercise:

Problem:

The force exerted by a diving board is conservative, provided the internal friction is negligible. Assuming friction is negligible, describe changes in the potential energy of a diving board as a swimmer dives from it, starting just before the swimmer steps on the board until just after his feet leave it.

Exercise:

Problem:

Define mechanical energy. What is the relationship of mechanical energy to nonconservative forces? What happens to mechanical energy if only conservative forces act?

Exercise:

Problem:

What is the relationship of potential energy to conservative force?

Problems & Exercises

Exercise:

Problem:

A 5.00×10^5-kg subway train is brought to a stop from a speed of 0.500 m/s in 0.400 m by a large spring bumper at the end of its track. What is the force constant k of the spring?

Solution:
Equation:

$$7.81 \times 10^5 \text{ N/m}$$

Exercise:

Problem:

A pogo stick has a spring with a force constant of $2.50 \times 10^4 \text{ N/m}$, which can be compressed 12.0 cm. To what maximum height can a child jump on the stick using only the energy in the spring, if the child and stick have a total mass of 40.0 kg? Explicitly show how you follow the steps in the Problem-Solving Strategies for Energy.

Glossary

conservative force
: a force that does the same work for any given initial and final configuration, regardless of the path followed

potential energy
: energy due to position, shape, or configuration

potential energy of a spring
: the stored energy of a spring as a function of its displacement; when Hooke's law applies, it is given by the expression $\frac{1}{2}kx^2$ where x is the distance the spring is compressed or extended and k is the spring constant

conservation of mechanical energy
: the rule that the sum of the kinetic energies and potential energies remains constant if only conservative forces act on and within a system

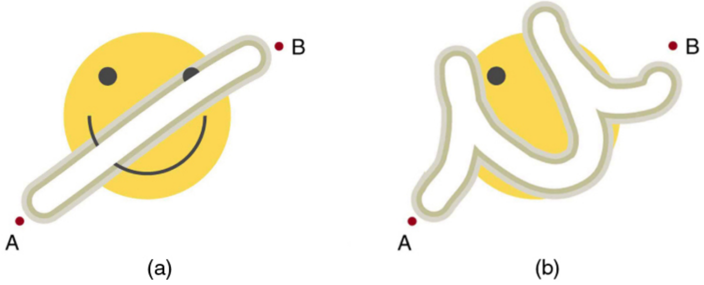
mechanical energy
: the sum of kinetic energy and potential energy

Nonconservative Forces

- Define nonconservative forces and explain how they affect mechanical energy.
- Show how the principle of conservation of energy can be applied by treating the conservative forces in terms of their potential energies and any nonconservative forces in terms of the work they do.

Nonconservative Forces and Friction

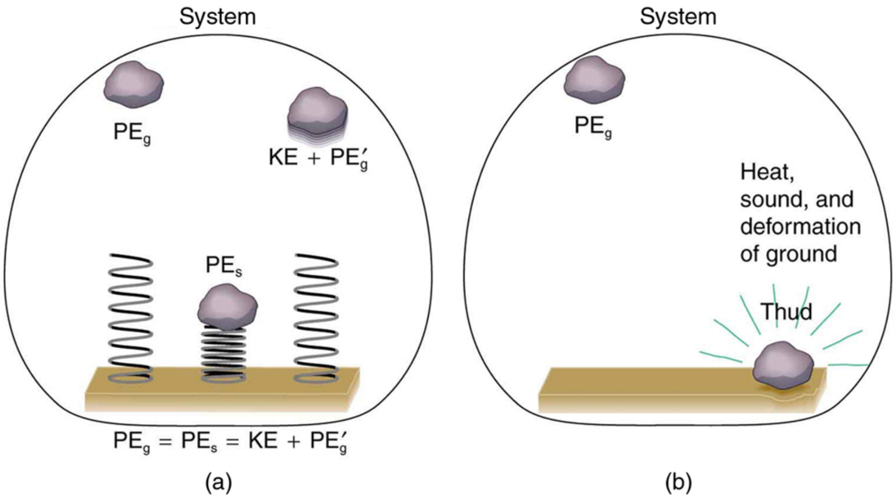
Forces are either conservative or nonconservative. Conservative forces were discussed in Conservative Forces and Potential Energy. A **nonconservative force** is one for which work depends on the path taken. Friction is a good example of a nonconservative force. As illustrated in [link], work done against friction depends on the length of the path between the starting and ending points. Because of this dependence on path, there is no potential energy associated with nonconservative forces. An important characteristic is that the work done by a nonconservative force *adds or removes mechanical energy from a system*. **Friction**, for example, creates **thermal energy** that dissipates, removing energy from the system. Furthermore, even if the thermal energy is retained or captured, it cannot be fully converted back to work, so it is lost or not recoverable in that sense as well.



The amount of the happy face erased depends on the path taken by the eraser between points A and B, as does the work done against friction. Less work is done and less of the face is erased for the path in (a) than for the path in (b). The force here is friction, and most of the work goes into thermal energy that subsequently leaves the system (the happy face plus the eraser). The energy expended cannot be fully recovered.

How Nonconservative Forces Affect Mechanical Energy

Mechanical energy *may* not be conserved when nonconservative forces act. For example, when a car is brought to a stop by friction on level ground, it loses kinetic energy, which is dissipated as thermal energy, reducing its mechanical energy. [link] compares the effects of conservative and nonconservative forces. We often choose to understand simpler systems such as that described in [link](a) first before studying more complicated systems as in [link](b).


System
PE$_g$
KE + PE$'_g$
PE$_s$
PE$_g$ = PE$_s$ = KE + PE$'_g$
(a)
System
PE$_g$
Heat, sound, and deformation of ground
Thud
(b)

Comparison of the effects of conservative and nonconservative forces on the mechanical energy of a system. (a) A system with only conservative forces. When a rock is dropped onto a spring, its mechanical energy remains constant (neglecting air resistance) because the force in the spring is conservative. The spring can propel the rock back to its original height, where it once again has only potential energy due to gravity. (b) A system with nonconservative forces. When the same rock is dropped onto the ground, it is stopped by nonconservative forces that dissipate its mechanical energy as thermal energy, sound, and surface distortion. The rock has lost mechanical energy.

How the Work-Energy Theorem Applies

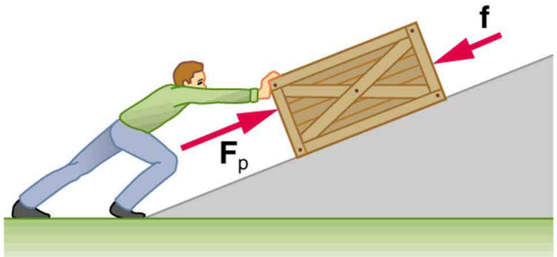
Now let us consider what form the work-energy theorem takes when both conservative and nonconservative forces act. We will see that the work done by nonconservative forces equals the change in the mechanical energy of a system. As noted in Kinetic Energy and the Work-Energy Theorem, the work-energy theorem states that the net work on a system equals the change in its kinetic energy, or $W_{\text{net}} = \Delta\text{KE}$. The net work is the sum of the work by nonconservative forces plus the work by conservative forces. That is,
Equation:

$$W_{\text{net}} = W_{\text{nc}} + W_{\text{c}},$$

so that
Equation:

$$W_{\text{nc}} + W_{\text{c}} = \Delta\text{KE},$$

where W_{nc} is the total work done by all nonconservative forces and W_{c} is the total work done by all conservative forces.



A person pushes a crate up a ramp, doing work on the crate. Friction and gravitational force (not shown) also do work on the crate; both forces oppose the person’s push. As the crate is pushed up the ramp, it gains mechanical energy, implying that the work done by the person is greater than the work done by friction.

Consider [link], in which a person pushes a crate up a ramp and is opposed by friction. As in the previous section, we note that work done by a conservative force comes from a loss of gravitational potential energy, so that $W_{\text{c}} = -\Delta\text{PE}$. Substituting this equation into the previous one and solving for W_{nc} gives

Equation:

$$W_{\text{nc}} = \Delta\text{KE} + \Delta\text{PE}.$$

This equation means that the total mechanical energy $(\text{KE} + \text{PE})$ changes by exactly the amount of work done by nonconservative forces. In [link], this is the work done by the person minus the work done by friction. So even if energy is not conserved for the system of interest (such as the crate), we know that an equal amount of work was done to cause the change in total mechanical energy.

We rearrange $W_{\text{nc}} = \Delta\text{KE} + \Delta\text{PE}$ to obtain
Equation:

$$\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}.$$

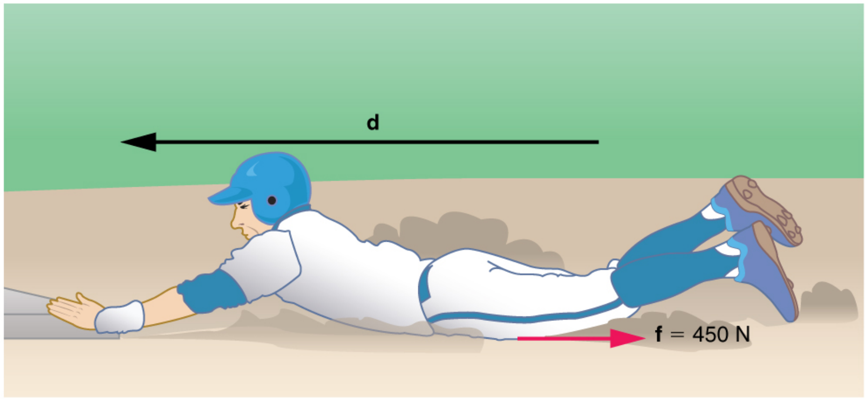
This means that the amount of work done by nonconservative forces adds to the mechanical energy of a system. If W_{nc} is positive, then mechanical energy is increased, such as when the person pushes the crate up the ramp in [link]. If W_{nc} is negative, then mechanical energy is decreased, such as when the rock hits the ground in [link](b). If W_{nc} is zero, then mechanical energy is conserved, and nonconservative forces are balanced. For example, when you push a lawn mower at constant speed on level ground, your work done is removed by the work of friction, and the mower has a constant energy.

Applying Energy Conservation with Nonconservative Forces

When no change in potential energy occurs, applying $\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}$ amounts to applying the work-energy theorem by setting the change in kinetic energy to be equal to the net work done on the system, which in the most general case includes both conservative and nonconservative forces. But when seeking instead to find a change in total mechanical energy in situations that involve changes in both potential and kinetic energy, the previous equation $\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}$ says that you can start by finding the change in mechanical energy that would have resulted from just the conservative forces, including the potential energy changes, and add to it the work done, with the proper sign, by any nonconservative forces involved.

Example:
Calculating Distance Traveled: How Far a Baseball Player Slides
Consider the situation shown in [link], where a baseball player slides to a stop on level ground. Using energy considerations, calculate the distance

the 65.0-kg baseball player slides, given that his initial speed is 6.00 m/s and the force of friction against him is a constant 450 N.



The baseball player slides to a stop in a distance *d*. In the process, friction removes the player’s kinetic energy by doing an amount of work fd equal to the initial kinetic energy.

Strategy

Friction stops the player by converting his kinetic energy into other forms, including thermal energy. In terms of the work-energy theorem, the work done by friction, which is negative, is added to the initial kinetic energy to reduce it to zero. The work done by friction is negative, because **f** is in the opposite direction of the motion (that is, $\theta = 180^{\mathrm{o}}$, and so $\cos\,\theta = -1$). Thus $W_{\mathrm{nc}} = -\mathrm{fd}$. The equation simplifies to

Equation:

$$\frac{1}{2}m{v_{\mathrm{i}}}^{2} - \mathrm{fd} = 0$$

or

Equation:

$$\mathrm{fd} = \frac{1}{2}m{v_{\mathrm{i}}}^{2}.$$

This equation can now be solved for the distance *d*.

Solution

Solving the previous equation for d and substituting known values yields

Equation:

$$\begin{aligned} d &= \frac{m{v_\mathrm{i}}^2}{2f} \\ &= \frac{(65.0\ \mathrm{kg})(6.00\ \mathrm{m/s})^2}{(2)(450\ \mathrm{N})} \\ &= 2.60\ \mathrm{m}. \end{aligned}$$

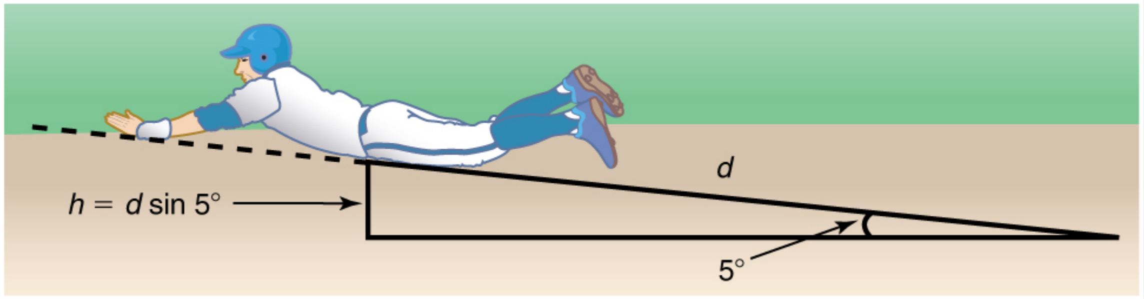
Discussion

The most important point of this example is that the amount of nonconservative work equals the change in mechanical energy. For example, you must work harder to stop a truck, with its large mechanical energy, than to stop a mosquito.

Example:

Calculating Distance Traveled: Sliding Up an Incline

Suppose that the player from [link] is running up a hill having a 5.00º incline upward with a surface similar to that in the baseball stadium. The player slides with the same initial speed, and the frictional force is still 450 N. Determine how far he slides.



The same baseball player slides to a stop on a 5.00º slope.

Strategy

In this case, the work done by the nonconservative friction force on the player reduces the mechanical energy he has from his kinetic energy at zero height, to the final mechanical energy he has by moving through distance d to reach height h along the hill, with $h = d \sin 5.00^{\circ}$. This is expressed by the equation

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} + W_{\mathrm{nc}} = \mathrm{KE_f} + \mathrm{PE_f}.$$

Solution

The work done by friction is again $W_{\mathrm{nc}} = -\mathrm{fd}$; initially the potential energy is $\mathrm{PE_i} = \mathrm{mg} \cdot 0 = 0$ and the kinetic energy is $\mathrm{KE_i} = \frac{1}{2} m {v_{\mathrm{i}}}^2$; the final energy contributions are $\mathrm{KE_f} = 0$ for the kinetic energy and $\mathrm{PE_f} = \mathrm{mgh} = \mathrm{mgd}\ \sin\theta$ for the potential energy.

Substituting these values gives

Equation:

$$\frac{1}{2} m {v_{\mathrm{i}}}^2 + 0 + \left(-fd\right) = 0 + mgd \sin\theta.$$

Solve this for d to obtain

Equation:

$$\begin{aligned} d &= \frac{\left(\frac{1}{2}\right) m {v_{\mathrm{i}}}^2}{f + mg \sin\theta} \\ &= \frac{(0.5)(65.0\ \mathrm{kg})(6.00\ \mathrm{m/s})^2}{450\ \mathrm{N} + (65.0\ \mathrm{kg})(9.80\ \mathrm{m/s}^2)\ \sin\ (5.00^{\circ})} \\ &= 2.31\ \mathrm{m}. \end{aligned}$$

Discussion

As might have been expected, the player slides a shorter distance by sliding uphill. Note that the problem could also have been solved in terms of the forces directly and the work energy theorem, instead of using the potential energy. This method would have required combining the normal force and force of gravity vectors, which no longer cancel each other because they point in different directions, and friction, to find the net force. You could then use the net force and the net work to find the distance d that reduces the kinetic energy to zero. By applying conservation of energy and using the potential energy instead, we need only consider the gravitational potential energy mgh, without combining and resolving force vectors. This simplifies the solution considerably.

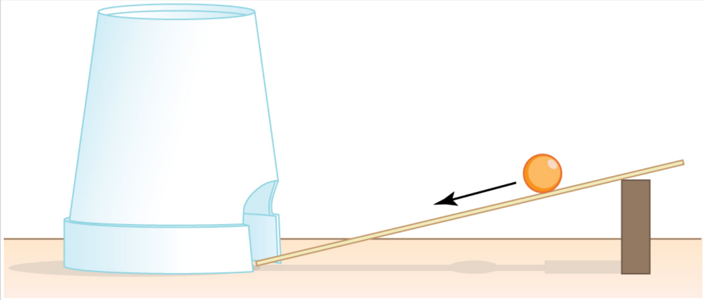
Note:

Making Connections: Take-Home Investigation—Determining Friction from the Stopping Distance

This experiment involves the conversion of gravitational potential energy into thermal energy. Use the ruler, book, and marble from Take-Home Investigation—Converting Potential to Kinetic Energy. In addition, you will need a foam cup with a small hole in the side, as shown in [link]. From the 10-cm position on the ruler, let the marble roll into the cup positioned at the bottom of the ruler. Measure the distance d the cup moves before stopping. What forces caused it to stop? What happened to the kinetic energy of the marble at the bottom of the ruler? Next, place the marble at the 20-cm and the 30-cm positions and again measure the distance the cup moves after the marble enters it. Plot the distance the cup moves versus the initial marble position on the ruler. Is this relationship linear?

With some simple assumptions, you can use these data to find the coefficient of kinetic friction μ_k of the cup on the table. The force of friction f on the cup is $\mu_k N$, where the normal force N is just the weight of the cup plus the marble. The normal force and force of gravity do no work because they are perpendicular to the displacement of the cup, which moves horizontally. The work done by friction is fd. You will need the mass of the marble as well to calculate its initial kinetic energy.

It is interesting to do the above experiment also with a steel marble (or ball bearing). Releasing it from the same positions on the ruler as you did with the glass marble, is the velocity of this steel marble the same as the velocity of the marble at the bottom of the ruler? Is the distance the cup moves proportional to the mass of the steel and glass marbles?



Rolling a marble down a ruler into a foam cup.

Note:

PhET Explorations: The Ramp

Explore forces, energy and work as you push household objects up and down a ramp. Lower and raise the ramp to see how the angle of inclination affects the parallel forces acting on the file cabinet. Graphs show forces, energy and work.

The Ramp

Section Summary

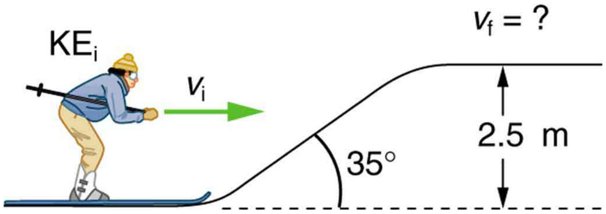
- A nonconservative force is one for which work depends on the path.
- Friction is an example of a nonconservative force that changes mechanical energy into thermal energy.
- Work W_{nc} done by a nonconservative force changes the mechanical energy of a system. In equation form, $W_{\text{nc}} = \Delta\text{KE} + \Delta\text{PE}$ or, equivalently, $\text{KE}_{\text{i}} + \text{PE}_{\text{i}} + W_{\text{nc}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}$.
- When both conservative and nonconservative forces act, energy conservation can be applied and used to calculate motion in terms of the known potential energies of the conservative forces and the work done by nonconservative forces, instead of finding the net work from the net force, or having to directly apply Newton’s laws.

Problems & Exercises

Exercise:

Problem:

A 60.0-kg skier with an initial speed of 12.0 m/s coasts up a 2.50-m-high rise as shown in [link]. Find her final speed at the top, given that the coefficient of friction between her skis and the snow is 0.0800. (Hint: Find the distance traveled up the incline assuming a straight-line path as shown in the figure.)



The skier’s initial kinetic energy is partially used in coasting to the top of a rise.

Solution:

9.46 m/s

Exercise:

Problem:

(a) How high a hill can a car coast up (engine disengaged) if work done by friction is negligible and its initial speed is 110 km/h? (b) If, in actuality, a 750-kg car with an initial speed of 110 km/h is observed to coast up a hill to a height 22.0 m above its starting point, how much thermal energy was generated by friction? (c) What is the average force of friction if the hill has a slope 2.5° above the horizontal?

Glossary

nonconservative force

a force whose work depends on the path followed between the given initial and final configurations

friction

the force between surfaces that opposes one sliding on the other; friction changes mechanical energy into thermal energy

Conservation of Energy

- Explain the law of the conservation of energy.
- Describe some of the many forms of energy.
- Define efficiency of an energy conversion process as the fraction left as useful energy or work, rather than being transformed, for example, into thermal energy.

Law of Conservation of Energy

Energy, as we have noted, is conserved, making it one of the most important physical quantities in nature. The **law of conservation of energy** can be stated as follows:

Total energy is constant in any process. It may change in form or be transferred from one system to another, but the total remains the same.

We have explored some forms of energy and some ways it can be transferred from one system to another. This exploration led to the definition of two major types of energy—mechanical energy $(\text{KE}+\text{PE})$ and energy transferred via work done by nonconservative forces (W_{nc}). But energy takes *many* other forms, manifesting itself in *many* different ways, and we need to be able to deal with all of these before we can write an equation for the above general statement of the conservation of energy.

Other Forms of Energy than Mechanical Energy

At this point, we deal with all other forms of energy by lumping them into a single group called other energy (OE). Then we can state the conservation of energy in equation form as
Equation:

$$\text{KE}_{\text{i}}+\text{PE}_{\text{i}}+W_{\text{nc}}+\text{OE}_{\text{i}}=\text{KE}_{\text{f}}+\text{PE}_{\text{f}}+\text{OE}_{\text{f}}.$$

All types of energy and work can be included in this very general statement of conservation of energy. Kinetic energy is KE, work done by a conservative force is represented by PE, work done by nonconservative forces is W_{nc}, and all other energies are included as OE. This equation applies to all previous examples; in those situations OE was constant, and so it subtracted out and was not directly considered.

Note:

Making Connections: Usefulness of the Energy Conservation Principle

The fact that energy is conserved and has many forms makes it very important. You will find that energy is discussed in many contexts, because it is involved in all processes. It will also become apparent that many situations are best understood in terms of energy and that problems are often most easily conceptualized and solved by considering energy.

When does OE play a role? One example occurs when a person eats. Food is oxidized with the release of carbon dioxide, water, and energy. Some of this chemical energy is converted to kinetic energy when the person moves, to potential energy when the person changes altitude, and to thermal energy (another form of OE).

Some of the Many Forms of Energy

What are some other forms of energy? You can probably name a number of forms of energy not yet discussed. Many of these will be covered in later chapters, but let us detail a few here. **Electrical energy** is a common form that is converted to many other forms and does work in a wide range of practical situations. Fuels, such as gasoline and food, carry **chemical energy** that can be transferred to a system through oxidation. Chemical fuel can also produce electrical energy, such as in batteries. Batteries can in turn produce light, which is a very pure form of energy. Most energy sources on Earth are in fact stored energy from the energy we receive from the Sun. We sometimes refer to this as **radiant energy**, or electromagnetic radiation, which includes visible light, infrared, and ultraviolet radiation. **Nuclear energy** comes from processes that convert measurable amounts of mass into energy. Nuclear energy is transformed into the energy of sunlight, into electrical energy in power plants, and into the energy of the heat transfer and blast in weapons.

Atoms and molecules inside all objects are in random motion. This internal mechanical energy from the random motions is called **thermal energy**, because it is related to the temperature of the object. These and all other forms of energy can be converted into one another and can do work.

[link] gives the amount of energy stored, used, or released from various objects and in various phenomena. The range of energies and the variety of types and situations is impressive.

Note:

Problem-Solving Strategies for Energy

You will find the following problem-solving strategies useful whenever you deal with energy. The strategies help in organizing and reinforcing energy concepts. In fact, they are used in the examples presented in this chapter. The familiar general problem-solving strategies presented earlier—involving identifying physical principles, knowns, and unknowns, checking units, and so on—continue to be relevant here.

Step 1. Determine the system of interest and identify what information is given and what quantity is to be calculated. A sketch will help.

Step 2. Examine all the forces involved and determine whether you know or are given the potential energy from the work done by the forces. Then use step 3 or step 4.

Step 3. If you know the potential energies for the forces that enter into the problem, then forces are all conservative, and you can apply conservation of mechanical energy simply in terms of potential and kinetic energy. The equation expressing conservation of energy is

Equation:

$$\text{KE}_{\text{i}} + \text{PE}_{\text{i}} = \text{KE}_{\text{f}} + \text{PE}_{\text{f}}.$$

Step 4. If you know the potential energy for only some of the forces, possibly because some of them are nonconservative and do not have a potential energy, or if there are other energies that are not easily treated in terms of force and work, then the conservation of energy law in its most general form must be used.

Equation:

$$\mathrm{KE_i} + \mathrm{PE_i} + W_{\mathrm{nc}} + \mathrm{OE_i} = \mathrm{KE_f} + \mathrm{PE_f} + \mathrm{OE_f}.$$

In most problems, one or more of the terms is zero, simplifying its solution. Do not calculate W_{c}, the work done by conservative forces; it is already incorporated in the PE terms.
Step 5. You have already identified the types of work and energy involved (in step 2). Before solving for the unknown, *eliminate terms wherever possible* to simplify the algebra. For example, choose $h = 0$ at either the initial or final point, so that $\mathrm{PE_g}$ is zero there. Then solve for the unknown in the customary manner.
Step 6. *Check the answer to see if it is reasonable*. Once you have solved a problem, reexamine the forms of work and energy to see if you have set up the conservation of energy equation correctly. For example, work done against friction should be negative, potential energy at the bottom of a hill should be less than that at the top, and so on. Also check to see that the numerical value obtained is reasonable. For example, the final speed of a skateboarder who coasts down a 3-m-high ramp could reasonably be 20 km/h, but *not* 80 km/h.

Transformation of Energy

The transformation of energy from one form into others is happening all the time. The chemical energy in food is converted into thermal energy through metabolism; light energy is converted into chemical energy through photosynthesis. In a larger example, the chemical energy contained in coal is converted into thermal energy as it burns to turn water into steam in a boiler. This thermal energy in the steam in turn is converted to mechanical energy as it spins a turbine, which is connected to a generator to produce electrical energy. (In all of these examples, not all of the initial energy is converted into the forms mentioned. This important point is discussed later in this section.)

Another example of energy conversion occurs in a solar cell. Sunlight impinging on a solar cell (see [link]) produces electricity, which in turn can be used to run an electric motor. Energy is converted from the primary source of solar energy into electrical energy and then into mechanical energy.

Solar energy is converted into electrical energy by solar cells, which is used to run a motor in this solar-power aircraft. (credit: NASA)

Object/phenomenon	**Energy in joules**
Big Bang	10^{68}
Energy released in a supernova	10^{44}
Fusion of all the hydrogen in Earth’s oceans	10^{34}
Annual world energy use	4×10^{20}

Object/phenomenon	**Energy in joules**
Large fusion bomb (9 megaton)	3.8×10^{16}
1 kg hydrogen (fusion to helium)	6.4×10^{14}
1 kg uranium (nuclear fission)	8.0×10^{13}
Hiroshima-size fission bomb (10 kiloton)	4.2×10^{13}
90,000-ton aircraft carrier at 30 knots	1.1×10^{10}
1 barrel crude oil	5.9×10^{9}
1 ton TNT	4.2×10^{9}
1 gallon of gasoline	1.2×10^{8}
Daily home electricity use (developed countries)	7×10^{7}
Daily adult food intake (recommended)	1.2×10^{7}

Object/phenomenon	Energy in joules
1000-kg car at 90 km/h	3.1×10^{5}
1 g fat (9.3 kcal)	3.9×10^{4}
ATP hydrolysis reaction	3.2×10^{4}
1 g carbohydrate (4.1 kcal)	1.7×10^{4}
1 g protein (4.1 kcal)	1.7×10^{4}
Tennis ball at 100 km/h	22
Mosquito $\left(10^{-2}\text{ g at }0.5\text{ m/s}\right)$	1.3×10^{-6}
Single electron in a TV tube beam	4.0×10^{-15}
Energy to break one DNA strand	10^{-19}

Energy of Various Objects and Phenomena

Efficiency

Even though energy is conserved in an energy conversion process, the output of *useful energy* or work will be less than the energy input. The **efficiency** Eff of an energy conversion process is defined as

Equation:

$$\text{Efficiency}(\text{Eff}) = \frac{\text{useful energy or work output}}{\text{total energy input}} = \frac{W_{\text{out}}}{E_{\text{in}}}.$$

[link] lists some efficiencies of mechanical devices and human activities. In a coal-fired power plant, for example, about 40% of the chemical energy in the coal becomes useful electrical energy. The other 60% transforms into other (perhaps less useful) energy forms, such as thermal energy, which is then released to the environment through combustion gases and cooling towers.

Activity/device	**Efficiency (%)[footnote]** **Representative values**
Cycling and climbing	20
Swimming, surface	2
Swimming, submerged	4
Shoveling	3
Weightlifting	9
Steam engine	17
Gasoline engine	30

Activity/device	Efficiency (%)[footnote] Representative values
Diesel engine	35
Nuclear power plant	35
Coal power plant	42
Electric motor	98
Compact fluorescent light	20
Gas heater (residential)	90
Solar cell	10

Efficiency of the Human Body and Mechanical Devices

Note:
PhET Explorations: Masses and Springs
A realistic mass and spring laboratory. Hang masses from springs and adjust the spring stiffness and damping. You can even slow time. Transport the lab to different planets. A chart shows the kinetic, potential, and thermal energies for each spring.
https://phet.colorado.edu/sims/mass-spring-lab/mass-spring-lab_en.html

Section Summary

- The law of conservation of energy states that the total energy is constant in any process. Energy may change in form or be transferred from one system to another, but the total remains the same.
- When all forms of energy are considered, conservation of energy is written in equation form as

$\mathrm{KE_i} + \mathrm{PE_i} + W_{\mathrm{nc}} + \mathrm{OE_i} = \mathrm{KE_f} + \mathrm{PE_f} + \mathrm{OE_f}$, where OE is all **other forms of energy** besides mechanical energy.

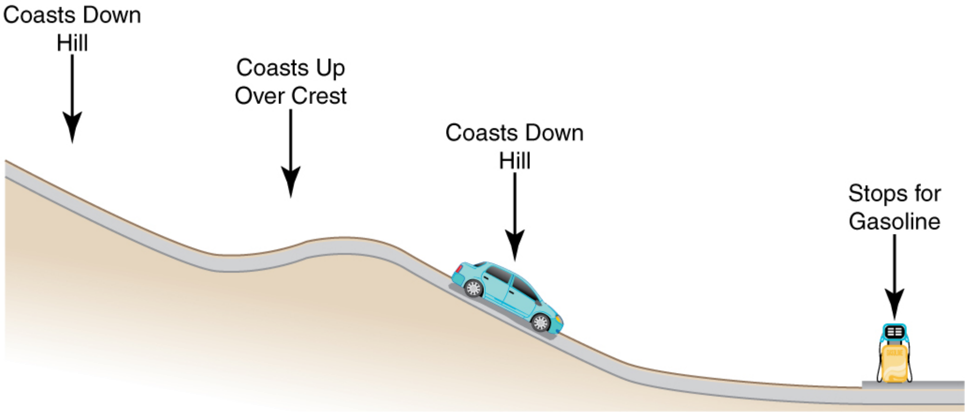
- Commonly encountered forms of energy include electric energy, chemical energy, radiant energy, nuclear energy, and thermal energy.
- Energy is often utilized to do work, but it is not possible to convert all the energy of a system to work.
- The efficiency Eff of a machine or human is defined to be $\mathrm{Eff} = \frac{W_{\mathrm{out}}}{E_{\mathrm{in}}}$, where W_{out} is useful work output and E_{in} is the energy consumed.

Conceptual Questions

Exercise:

Problem:

Consider the following scenario. A car for which friction is *not* negligible accelerates from rest down a hill, running out of gasoline after a short distance. The driver lets the car coast farther down the hill, then up and over a small crest. He then coasts down that hill into a gas station, where he brakes to a stop and fills the tank with gasoline. Identify the forms of energy the car has, and how they are changed and transferred in this series of events. (See [link].)



A car experiencing non-negligible friction coasts down a hill, over a small crest, then downhill again, and comes to a stop at a gas station.

Exercise:

Problem:

Describe the energy transfers and transformations for a javelin, starting from the point at which an athlete picks up the javelin and ending when the javelin is stuck into the ground after being thrown.

Exercise:

Problem:

Do devices with efficiencies of less than one violate the law of conservation of energy? Explain.

Exercise:

Problem:

List four different forms or types of energy. Give one example of a conversion from each of these forms to another form.

Exercise:

Problem: List the energy conversions that occur when riding a bicycle.

Problems & Exercises

Exercise:

Problem:

Using values from [link], how many DNA molecules could be broken by the energy carried by a single electron in the beam of an old-fashioned TV tube? (These electrons were not dangerous in themselves, but they did create dangerous x rays. Later model tube TVs had shielding that absorbed x rays before they escaped and exposed viewers.)

Solution:

4×10^4 molecules

Exercise:

Problem:

Using energy considerations and assuming negligible air resistance, show that a rock thrown from a bridge 20.0 m above water with an initial speed of 15.0 m/s strikes the water with a speed of 24.8 m/s independent of the direction thrown.

Solution:

Equating ΔPE_g and ΔKE, we obtain

$$v=\sqrt{2\text{gh}+{v_0}^2}=\sqrt{2(9.80\text{ m/s}^2)(20.0\text{ m})+(15.0\text{ m/s})^2}=24.8\text{ m/s}$$

Exercise:

Problem:

If the energy in fusion bombs were used to supply the energy needs of the world, how many of the 9-megaton variety would be needed for a year's supply of energy (using data from [link])? This is not as far-fetched as it may sound—there are thousands of nuclear bombs, and their energy can be trapped in underground explosions and converted to electricity, as natural geothermal energy is.

Exercise:

Problem:

(a) Use of hydrogen fusion to supply energy is a dream that may be realized in the next century. Fusion would be a relatively clean and almost limitless supply of energy, as can be seen from [link]. To illustrate this, calculate how many years the present energy needs of the world could be supplied by one millionth of the oceans' hydrogen fusion energy. (b) How does this time compare with historically significant events, such as the duration of stable economic systems?

Solution:

(a) 25×10^6 years

(b) This is much, much longer than human time scales.

Glossary

law of conservation of energy
the general law that total energy is constant in any process; energy may change in form or be transferred from one system to another, but the total remains the same

electrical energy
the energy carried by a flow of charge

chemical energy
the energy in a substance stored in the bonds between atoms and molecules that can be released in a chemical reaction

radiant energy
the energy carried by electromagnetic waves

nuclear energy
energy released by changes within atomic nuclei, such as the fusion of two light nuclei or the fission of a heavy nucleus

thermal energy
the energy within an object due to the random motion of its atoms and molecules that accounts for the object's temperature

efficiency
a measure of the effectiveness of the input of energy to do work; useful energy or work divided by the total input of energy

Power

- Calculate power by calculating changes in energy over time.
- Examine power consumption and calculations of the cost of energy consumed.

What is Power?

Power—the word conjures up many images: a professional football player muscling aside his opponent, a dragster roaring away from the starting line, a volcano blowing its lava into the atmosphere, or a rocket blasting off, as in [link].

This powerful rocket on the Space Shuttle *Endeavor* did work and consumed energy at a very high rate. (credit: NASA)

These images of power have in common the rapid performance of work, consistent with the scientific definition of **power** (P) as the rate at which work is done.

Note:
Power
Power is the rate at which work is done.
Equation:

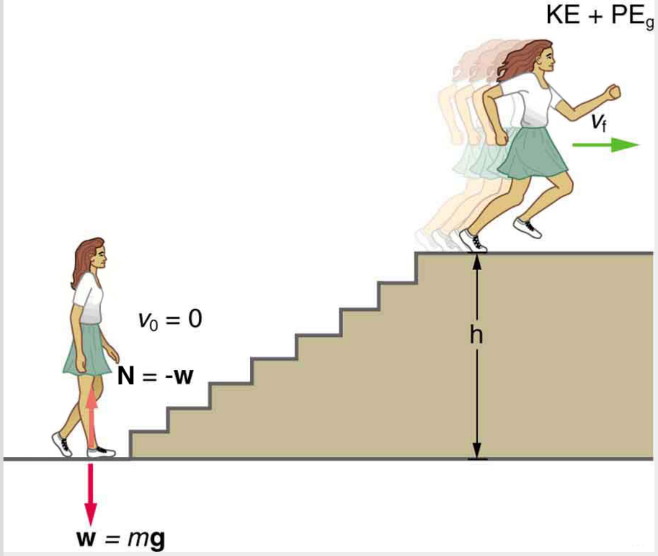
$$P=\frac{W}{t}$$

The SI unit for power is the **watt** (W), where 1 watt equals 1 joule/second $(1\text{ W}=1\text{ J/s})$.

Because work is energy transfer, power is also the rate at which energy is expended. A 60-W light bulb, for example, expends 60 J of energy per second. Great power means a large amount of work or energy developed in a short time. For example, when a powerful car accelerates rapidly, it does a large amount of work and consumes a large amount of fuel in a short time.

Calculating Power from Energy

Example:
Calculating the Power to Climb Stairs
What is the power output for a 60.0-kg woman who runs up a 3.00 m high flight of stairs in 3.50 s, starting from rest but having a final speed of 2.00 m/s? (See [link].)



When this woman runs upstairs starting from rest, she converts the chemical energy originally from food into kinetic energy and gravitational potential energy. Her power output depends on how fast she does this.

Strategy and Concept

The work going into mechanical energy is $W = \text{KE} + \text{PE}$. At the bottom of the stairs, we take both KE and PE_g as initially zero; thus, $W = \text{KE}_\text{f} + \text{PE}_\text{g} = \frac{1}{2}m{v_\text{f}}^2 + mgh$, where h is the vertical height of the stairs. Because all terms are given, we can calculate W and then divide it by time to get power.

Solution

Substituting the expression for W into the definition of power given in the previous equation, $P = W/t$ yields

Equation:

$$P = \frac{W}{t} = \frac{\frac{1}{2}m{v_\text{f}}^2 + mgh}{t}.$$

Entering known values yields

Equation:

$$\begin{aligned} P &= \frac{0.5(60.0\text{ kg})(2.00\text{ m/s})^2 + (60.0\text{ kg})\left(9.80\text{ m/s}^2\right)(3.00\text{ m})}{3.50\text{ s}} \\ &= \frac{120\text{ J} + 1764\text{ J}}{3.50\text{ s}} \\ &= 538\text{ W}. \end{aligned}$$

Discussion

The woman does 1764 J of work to move up the stairs compared with only 120 J to increase her kinetic energy; thus, most of her power output is required for climbing rather than accelerating.

It is impressive that this woman's useful power output is slightly less than 1 **horsepower** $(1\text{ hp} = 746\text{ W})$! People can generate more than a horsepower with their leg muscles for short periods of time by rapidly converting available blood sugar and oxygen into work output. (A horse can put out 1 hp for hours on end.) Once oxygen is depleted, power output decreases and the person begins to breathe rapidly to obtain oxygen to metabolize more food—this is known as the *aerobic* stage of exercise. If the woman climbed the stairs slowly, then her power output would be much less, although the amount of work done would be the same.

Note:

Making Connections: Take-Home Investigation—Measure Your Power Rating

Determine your own power rating by measuring the time it takes you to climb a flight of stairs. We will ignore the gain in kinetic energy, as the above example showed that it was a small portion of the energy gain. Don't expect that your output will be more than about 0.5 hp.

Examples of Power

Examples of power are limited only by the imagination, because there are as many types as there are forms of work and energy. (See [link] for some examples.) Sunlight reaching Earth’s surface carries a maximum power of about 1.3 kilowatts per square meter (kW/m^2). A tiny fraction of this is retained by Earth over the long term. Our consumption rate of fossil fuels is far greater than the rate at which they are stored, so it is inevitable that they will be depleted. Power implies that energy is transferred, perhaps changing form. It is never possible to change one form completely into another without losing some of it as thermal energy. For example, a 60-W incandescent bulb converts only 5 W of electrical power to light, with 55 W dissipating into thermal energy. Furthermore, the typical electric power plant converts only 35 to 40% of its fuel into electricity. The remainder becomes a huge amount of thermal energy that must be dispersed as heat transfer, as rapidly as it is created. A coal-fired power plant may produce 1000 megawatts; 1 megawatt (MW) is 10^6 W of electric power. But the power plant consumes chemical energy at a rate of about 2500 MW, creating heat transfer to the surroundings at a rate of 1500 MW. (See [link].)

Tremendous amounts of electric power are generated by coal-fired power plants such as this one in China, but an even larger amount of power goes into heat transfer to the surroundings.

The large cooling towers here are needed to transfer heat as rapidly as it is produced. The transfer of heat is not unique to coal plants but is an unavoidable consequence of generating electric power from any fuel—nuclear, coal, oil, natural gas, or the like. (credit: Kleinolive, Wikimedia Commons)

Object or Phenomenon	**Power in Watts**
Supernova (at peak)	5×10^{37}
Milky Way galaxy	10^{37}
Crab Nebula pulsar	10^{28}
The Sun	4×10^{26}

Object or Phenomenon	**Power in Watts**
Volcanic eruption (maximum)	4×10^{15}
Lightning bolt	2×10^{12}
Nuclear power plant (total electric and heat transfer)	3×10^{9}
Aircraft carrier (total useful and heat transfer)	10^{8}
Dragster (total useful and heat transfer)	2×10^{6}
Car (total useful and heat transfer)	8×10^{4}
Football player (total useful and heat transfer)	5×10^{3}
Clothes dryer	4×10^{3}
Person at rest (all heat transfer)	100

Object or Phenomenon	Power in Watts
Typical incandescent light bulb (total useful and heat transfer)	60
Heart, person at rest (total useful and heat transfer)	8
Electric clock	3
Pocket calculator	10^{-3}

Power Output or Consumption

Power and Energy Consumption

We usually have to pay for the energy we use. It is interesting and easy to estimate the cost of energy for an electrical appliance if its power consumption rate and time used are known. The higher the power consumption rate and the longer the appliance is used, the greater the cost of that appliance. The power consumption rate is $P=W/t=E/t$, where E is the energy supplied by the electricity company. So the energy consumed over a time t is

Equation:

$$E=\text{Pt}.$$

Electricity bills state the energy used in units of **kilowatt-hours** $(\text{kW}\cdot\text{h})$, which is the product of power in kilowatts and time in hours. This unit is convenient because electrical power consumption at the kilowatt level for hours at a time is typical.

Example:
Calculating Energy Costs

What is the cost of running a 0.200-kW computer 6.00 h per day for 30.0 d if the cost of electricity is $0.120 per $\mathrm{kW} \cdot \mathrm{h}$?

Strategy

Cost is based on energy consumed; thus, we must find E from $E = \mathrm{Pt}$ and then calculate the cost. Because electrical energy is expressed in $\mathrm{kW} \cdot \mathrm{h}$, at the start of a problem such as this it is convenient to convert the units into kW and hours.

Solution

The energy consumed in $\mathrm{kW} \cdot \mathrm{h}$ is

Equation:

$$\begin{aligned} E &= \mathrm{Pt} = (0.200\,\mathrm{kW})(6.00\,\mathrm{h/d})(30.0\,\mathrm{d}) \\ &= 36.0\,\mathrm{kW} \cdot \mathrm{h}, \end{aligned}$$

and the cost is simply given by

Equation:

$$\text{cost} = (36.0\ \mathrm{kW} \cdot \mathrm{h})(\$0.120\ \text{per}\ \mathrm{kW} \cdot \mathrm{h}) = \$4.32\ \text{per month.}$$

Discussion

The cost of using the computer in this example is neither exorbitant nor negligible. It is clear that the cost is a combination of power and time. When both are high, such as for an air conditioner in the summer, the cost is high.

The motivation to save energy has become more compelling with its ever-increasing price. Armed with the knowledge that energy consumed is the product of power and time, you can estimate costs for yourself and make the necessary value judgments about where to save energy. Either power or time must be reduced. It is most cost-effective to limit the use of high-power devices that normally operate for long periods of time, such as water heaters and air conditioners. This would not include relatively high power devices like toasters, because they are on only a few minutes per day. It would also not include electric clocks, in spite of their 24-hour-per-day usage, because they are very low power devices. It is sometimes possible to use devices that have greater efficiencies—that is, devices that consume less power to accomplish the same task. One example is the compact fluorescent light bulb, which produces over four times more light per watt of power consumed than its incandescent cousin.

Modern civilization depends on energy, but current levels of energy consumption and production are not sustainable. The likelihood of a link between global warming and fossil fuel use (with its concomitant production of carbon dioxide), has made reduction in energy use as well as a shift to non-fossil fuels of the utmost importance. Even though energy in an isolated system is a conserved quantity, the final result of most energy transformations is waste heat transfer to the environment, which is no longer useful for doing work. As we will discuss in more detail in Thermodynamics, the potential for energy to produce useful work has been "degraded" in the energy transformation.

Section Summary

- Power is the rate at which work is done, or in equation form, for the average power P for work W done over a time t, $P=W/t$.
- The SI unit for power is the watt (W), where $1\text{ W}=1\text{ J/s}$.
- The power of many devices such as electric motors is also often expressed in horsepower (hp), where $1\text{ hp}=746\text{ W}$.

Conceptual Questions

Exercise:

Problem:

Most electrical appliances are rated in watts. Does this rating depend on how long the appliance is on? (When off, it is a zero-watt device.) Explain in terms of the definition of power.

Exercise:

Problem:

Explain, in terms of the definition of power, why energy consumption is sometimes listed in kilowatt-hours rather than joules. What is the relationship between these two energy units?

Exercise:

Problem:

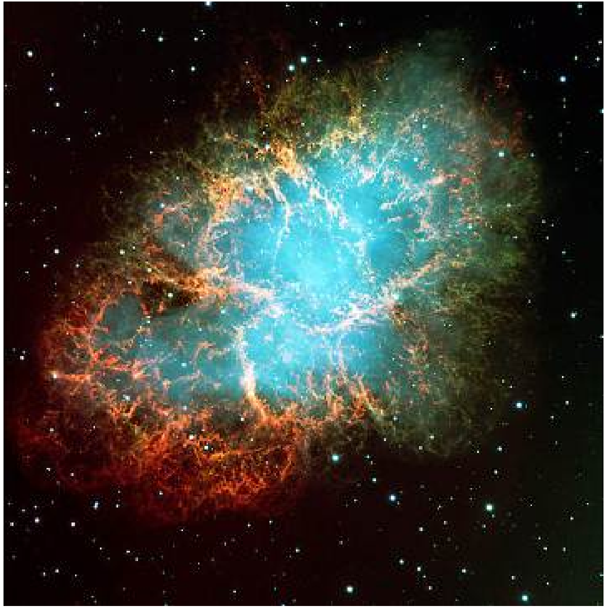
A spark of static electricity, such as that you might receive from a doorknob on a cold dry day, may carry a few hundred watts of power. Explain why you are not injured by such a spark.

Problems & Exercises

Exercise:

Problem:

The Crab Nebula (see [link]) pulsar is the remnant of a supernova that occurred in A.D. 1054. Using data from [link], calculate the approximate factor by which the power output of this astronomical object has declined since its explosion.



Crab Nebula (credit: ESO, via Wikimedia Commons)

Solution:
Equation:

$$2\times 10^{-10}$$

Exercise:

Problem:

Suppose a star 1000 times brighter than our Sun (that is, emitting 1000 times the power) suddenly goes supernova. Using data from [link]: (a) By what factor does its power output increase? (b) How many times brighter than our entire Milky Way galaxy is the supernova? (c) Based on your answers, discuss whether it should be possible to observe supernovas in distant galaxies. Note that there are on the order of 10^{11} observable galaxies, the average brightness of which is somewhat less than our own galaxy.

Exercise:

Problem:

A person in good physical condition can put out 100 W of useful power for several hours at a stretch, perhaps by pedaling a mechanism that drives an electric generator. Neglecting any problems of generator efficiency and practical considerations such as resting time: (a) How many people would it take to run a 4.00-kW electric clothes dryer? (b) How many people would it take to replace a large electric power plant that generates 800 MW?

Solution:

(a) 40

(b) 8 million

Exercise:

Problem:

What is the cost of operating a 3.00-W electric clock for a year if the cost of electricity is $0.0900 per $\mathrm{kW \cdot h}$?

Exercise:

Problem:

A large household air conditioner may consume 15.0 kW of power. What is the cost of operating this air conditioner 3.00 h per day for 30.0 d if the cost of electricity is $0.110 per $\mathrm{kW \cdot h}$?

Solution:

$149

Exercise:

Problem:

(a) What is the average power consumption in watts of an appliance that uses $5.00\ \mathrm{kW \cdot h}$ of energy per day? (b) How many joules of energy does this appliance consume in a year?

Exercise:

Problem:

(a) What is the average useful power output of a person who does 6.00×10^6 J of useful work in 8.00 h? (b) Working at this rate, how long will it take this person to lift 2000 kg of bricks 1.50 m to a platform? (Work done to lift his body can be omitted because it is not considered useful output here.)

Solution:

(a) $208\ \mathrm{W}$

(b) 141 s

Exercise:

Problem:

A 500-kg dragster accelerates from rest to a final speed of 110 m/s in 400 m (about a quarter of a mile) and encounters an average frictional force of 1200 N. What is its average power output in watts and horsepower if this takes 7.30 s?

Exercise:

Problem:

(a) How long will it take an 850-kg car with a useful power output of 40.0 hp (1 hp = 746 W) to reach a speed of 15.0 m/s, neglecting friction? (b) How long will this acceleration take if the car also climbs a 3.00-m-high hill in the process?

Solution:

(a) 3.20 s

(b) 4.04 s

Exercise:

Problem:

(a) Find the useful power output of an elevator motor that lifts a 2500-kg load a height of 35.0 m in 12.0 s, if it also increases the speed from rest to 4.00 m/s. Note that the total mass of the counterbalanced system is 10,000 kg—so that only 2500 kg is raised in height, but the full 10,000 kg is accelerated. (b) What does it cost, if electricity is $0.0900 per $\text{kW}\cdot\text{h}$?

Exercise:

Problem:

(a) What is the available energy content, in joules, of a battery that operates a 2.00-W electric clock for 18 months? (b) How long can a battery that can supply 8.00×10^4 J run a pocket calculator that consumes energy at the rate of 1.00×10^{-3} W?

Solution:

(a) 9.46×10^7 J

(b) 2.54 y

Exercise:

Problem:

(a) How long would it take a 1.50×10^5-kg airplane with engines that produce 100 MW of power to reach a speed of 250 m/s and an altitude of 12.0 km if air resistance were negligible? (b) If it actually takes 900 s, what is the power? (c) Given this power, what is the average force of air resistance if the airplane takes 1200 s? (Hint: You must find the distance the plane travels in 1200 s assuming constant acceleration.)

Exercise:

Problem:

Calculate the power output needed for a 950-kg car to climb a 2.00º slope at a constant 30.0 m/s while encountering wind resistance and friction totaling 600 N. Explicitly show how you follow the steps in the Problem-Solving Strategies for Energy.

Solution:

Identify knowns: $m = 950\text{ kg}$, slope angle $\theta = 2.00^\text{o}$, $v = 3.00\text{ m/s}$, $f = 600\text{ N}$

Identify unknowns: power P of the car, force F that car applies to road

Solve for unknown:

$$P=\frac{W}{t}=\frac{\text{Fd}}{t}=F\left(\frac{d}{t}\right)=\text{Fv},$$

where F is parallel to the incline and must oppose the resistive forces and the force of gravity:

$$F=f+w=600\text{ N}+\text{mg sin }\theta$$

Insert this into the expression for power and solve:

$$\begin{aligned} P &= (f+\text{mg sin }\theta)v \\ &= \left[600\text{ N}+(950\text{ kg})\left(9.80\text{ m/s}^2\right)\text{sin }2^{\text{o}}\right](30.0\text{ m/s}) \\ &= 2.77\times10^4\text{ W} \end{aligned}$$

About 28 kW (or about 37 hp) is reasonable for a car to climb a gentle incline.

Exercise:

Problem:

(a) Calculate the power per square meter reaching Earth’s upper atmosphere from the Sun. (Take the power output of the Sun to be 4.00×10^{26} W.) (b) Part of this is absorbed and reflected by the atmosphere, so that a maximum of 1.30 kW/m^2 reaches Earth’s surface. Calculate the area in km^2 of solar energy collectors needed to replace an electric power plant that generates 750 MW if the collectors convert an average of 2.00% of the maximum power into electricity. (This small conversion efficiency is due to the devices themselves, and the fact that the sun is directly overhead only briefly.) With the same assumptions, what area would be needed to meet the United States’ energy needs $(1.05\times10^{20}\text{ J})$? Australia’s energy needs $(5.4\times10^{18}\text{ J})$? China’s energy needs $(6.3\times10^{19}\text{ J})$? (These energy consumption values are from 2006.)

Glossary

power
: the rate at which work is done

watt
: (W) SI unit of power, with $1\text{ W}=1\text{ J/s}$

horsepower
: an older non-SI unit of power, with $1\text{ hp}=746\text{ W}$

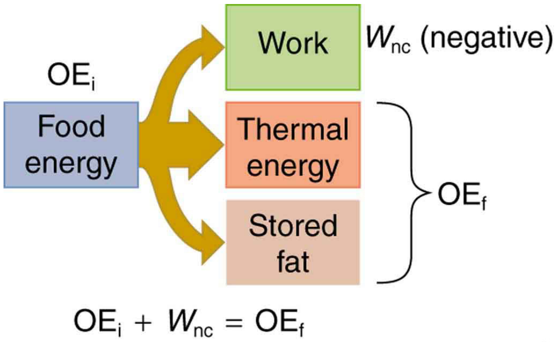
kilowatt-hour
: $(\text{kW}\cdot\text{h})$ unit used primarily for electrical energy provided by electric utility companies

Work, Energy, and Power in Humans

- Explain the human body’s consumption of energy when at rest vs. when engaged in activities that do useful work.
- Calculate the conversion of chemical energy in food into useful work.

Energy Conversion in Humans

Our own bodies, like all living organisms, are energy conversion machines. Conservation of energy implies that the chemical energy stored in food is converted into work, thermal energy, and/or stored as chemical energy in fatty tissue. (See [link].) The fraction going into each form depends both on how much we eat and on our level of physical activity. If we eat more than is needed to do work and stay warm, the remainder goes into body fat.



Energy consumed by humans is converted to work, thermal energy, and stored fat. By far the largest fraction goes to thermal energy, although the fraction varies depending on the type of physical activity.

Power Consumed at Rest

The *rate* at which the body uses food energy to sustain life and to do different activities is called the **metabolic rate**. The total energy conversion rate of a person *at rest* is called the **basal metabolic rate** (BMR) and is divided among various systems in the body, as shown in [link]. The largest fraction goes to the liver and spleen, with the brain coming next. Of course, during vigorous exercise, the energy consumption of the skeletal muscles and heart increase markedly. About 75% of the calories burned in a day go into these basic functions. The BMR is a function of age, gender, total body weight, and amount of muscle mass (which burns more calories than body fat). Athletes have a greater BMR due to this last factor.

Organ	**Power consumed at rest (W)**	**Oxygen consumption (mL/min)**	**Percent of BMR**
Liver & spleen	23	67	27
Brain	16	47	19
Skeletal muscle	15	45	18
Kidney	9	26	10
Heart	6	17	7
Other	16	48	19
Totals	85 W	250 mL/min	100%

Basal Metabolic Rates (BMR)

Energy consumption is directly proportional to oxygen consumption because the digestive process is basically one of oxidizing food. We can measure the energy people use during various activities by measuring their oxygen use. (See [link].) Approximately 20 kJ of energy are produced for each liter of oxygen consumed, independent of the type of food. [link] shows energy and oxygen consumption rates (power expended) for a variety of activities.

Power of Doing Useful Work

Work done by a person is sometimes called **useful work**, which is *work done on the outside world*, such as lifting weights. Useful work requires a force exerted through a distance on the outside world, and so it excludes internal work, such as that done by the heart when pumping blood. Useful work does include that done in climbing stairs or accelerating to a full run, because these are accomplished by exerting forces on the outside world. Forces exerted by the body are nonconservative, so that they can change the mechanical energy ($\mathrm{KE} + \mathrm{PE}$) of the system worked upon, and this is often the goal. A baseball player throwing a ball, for example, increases both the ball’s kinetic and potential energy.

If a person needs more energy than they consume, such as when doing vigorous work, the body must draw upon the chemical energy stored in fat. So exercise can be helpful in losing fat. However, the amount of exercise needed to produce a loss in fat, or to burn off extra calories consumed that day, can be large, as [link] illustrates.

Example:
Calculating Weight Loss from Exercising
If a person who normally requires an average of 12,000 kJ (3000 kcal) of food energy per day consumes 13,000 kJ per day, he will steadily gain weight. How much bicycling per day is required to work off this extra 1000 kJ?

Solution

[link] states that 400 W are used when cycling at a moderate speed. The time required to work off 1000 kJ at this rate is then

Equation:

$$\text{Time} = \frac{\text{energy}}{\left(\frac{\text{energy}}{\text{time}}\right)} = \frac{1000\text{ kJ}}{400\text{ W}} = 2500\text{ s} = 42\text{ min.}$$

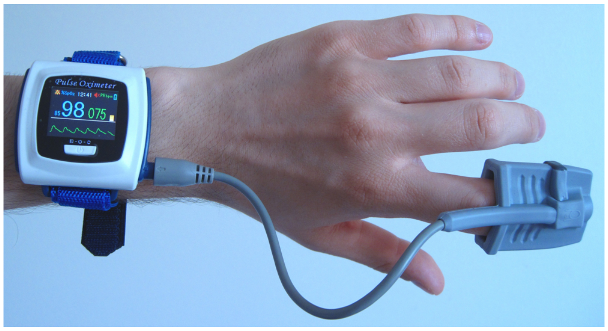
Discussion

If this person uses more energy than he or she consumes, the person’s body will obtain the needed energy by metabolizing body fat. If the person uses 13,000 kJ but consumes only 12,000 kJ, then the amount of fat loss will be

Equation:

$$\text{Fat loss} = (1000\text{ kJ})\left(\frac{1.0\text{ g fat}}{39\text{ kJ}}\right) = 26\text{ g},$$

assuming the energy content of fat to be 39 kJ/g.



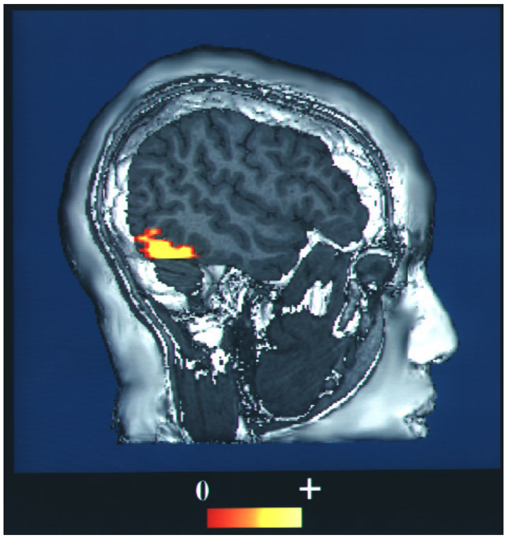
A pulse oxymeter is an apparatus that measures the amount of oxygen in blood. Oxymeters can be used to determine a person’s metabolic rate, which is the rate at which food energy is converted to another form. Such measurements can indicate the level of athletic conditioning as well as certain medical problems. (credit: UusiAjaja, Wikimedia Commons)

Activity	Energy consumption in watts	Oxygen consumption in liters O_2/min
Sleeping	83	0.24
Sitting at rest	120	0.34
Standing relaxed	125	0.36
Sitting in class	210	0.60
Walking (5 km/h)	280	0.80
Cycling (13–18 km/h)	400	1.14
Shivering	425	1.21
Playing tennis	440	1.26

Activity	Energy consumption in watts	Oxygen consumption in liters O_2/min
Swimming breaststroke	475	1.36
Ice skating (14.5 km/h)	545	1.56
Climbing stairs (116/min)	685	1.96
Cycling (21 km/h)	700	2.00
Running cross-country	740	2.12
Playing basketball	800	2.28
Cycling, professional racer	1855	5.30
Sprinting	2415	6.90

Energy and Oxygen Consumption Rates[footnote] (Power)
for an average 76-kg male

All bodily functions, from thinking to lifting weights, require energy. (See [link].) The many small muscle actions accompanying all quiet activity, from sleeping to head scratching, ultimately become thermal energy, as do less visible muscle actions by the heart, lungs, and digestive tract. Shivering, in fact, is an involuntary response to low body temperature that pits muscles against one another to produce thermal energy in the body (and do no work). The kidneys and liver consume a surprising amount of energy, but the biggest surprise of all is that a full 25% of all energy consumed by the body is used to maintain electrical potentials in all living cells. (Nerve cells use this electrical potential in nerve impulses.) This bioelectrical energy ultimately becomes mostly thermal energy, but some is utilized to power chemical processes such as in the kidneys and liver, and in fat production.



This fMRI scan shows an increased level of energy consumption in the vision center of the brain. Here, the patient was being asked to recognize faces. (credit: NIH via Wikimedia Commons)

Section Summary

- The human body converts energy stored in food into work, thermal energy, and/or chemical energy that is stored in fatty tissue.
- The *rate* at which the body uses food energy to sustain life and to do different activities is called the metabolic rate, and the corresponding rate when at rest is called the basal metabolic rate (BMR)
- The energy included in the basal metabolic rate is divided among various systems in the body, with the largest fraction going to the liver and spleen, and the brain coming next.
- About 75% of food calories are used to sustain basic body functions included in the basal metabolic rate.
- The energy consumption of people during various activities can be determined by measuring their oxygen use, because the digestive process is basically one of oxidizing food.

Conceptual Questions

Exercise:

Problem:

Explain why it is easier to climb a mountain on a zigzag path rather than one straight up the side. Is your increase in gravitational potential energy the same in both cases? Is your energy consumption the same in both?

Exercise:

Problem:

Do you do work on the outside world when you rub your hands together to warm them? What is the efficiency of this activity?

Exercise:

Problem:

Shivering is an involuntary response to lowered body temperature. What is the efficiency of the body when shivering, and is this a desirable value?

Exercise:

Problem:

Discuss the relative effectiveness of dieting and exercise in losing weight, noting that most athletic activities consume food energy at a rate of 400 to 500 W, while a single cup of yogurt can contain 1360 kJ (325 kcal). Specifically, is it likely that exercise alone will be sufficient to lose weight? You may wish to consider that regular exercise may increase the metabolic rate, whereas protracted dieting may reduce it.

Problems & Exercises

Exercise:

Problem:

(a) How long can you rapidly climb stairs (116/min) on the 93.0 kcal of energy in a 10.0-g pat of butter? (b) How many flights is this if each flight has 16 stairs?

Solution:

(a) 9.5 min

(b) 69 flights of stairs

Exercise:

Problem:

(a) What is the power output in watts and horsepower of a 70.0-kg sprinter who accelerates from rest to 10.0 m/s in 3.00 s? (b) Considering the amount of power generated, do you think a well-trained athlete could do this repetitively for long periods of time?

Exercise:

Problem:

Calculate the power output in watts and horsepower of a shot-putter who takes 1.20 s to accelerate the 7.27-kg shot from rest to 14.0 m/s, while raising it 0.800 m. (Do not include the power produced to accelerate his body.)

Shot putter at the Dornoch Highland Gathering in 2007. (credit: John Haslam, Flickr)

Solution:

641 W, 0.860 hp

Exercise:

Problem:

(a) What is the efficiency of an out-of-condition professor who does 2.10×10^{5} J of useful work while metabolizing 500 kcal of food energy? (b) How many food calories would a well-conditioned athlete metabolize in doing the same work with an efficiency of 20%?

Exercise:

Problem:

Energy that is not utilized for work or heat transfer is converted to the chemical energy of body fat containing about 39 kJ/g. How many grams of fat will you gain if you eat 10,000 kJ (about 2500 kcal) one day and do nothing but sit relaxed for 16.0 h and sleep for the other 8.00 h? Use data from [link] for the energy consumption rates of these activities.

Solution:

31 g

Exercise:

Problem:

Using data from [link], calculate the daily energy needs of a person who sleeps for 7.00 h, walks for 2.00 h, attends classes for 4.00 h, cycles for 2.00 h, sits relaxed for 3.00 h, and studies for 6.00 h. (Studying consumes energy at the same rate as sitting in class.)

Exercise:

Problem:

What is the efficiency of a subject on a treadmill who puts out work at the rate of 100 W while consuming oxygen at the rate of 2.00 L/min? (Hint: See [link].)

Solution:

14.3%

Exercise:

Problem:

Shoveling snow can be extremely taxing because the arms have such a low efficiency in this activity. Suppose a person shoveling a footpath metabolizes food at the rate of 800 W. (a) What is her useful power output? (b) How long will it take her to lift 3000 kg of snow 1.20 m? (This could be the amount of heavy snow on 20 m of footpath.) (c) How much waste heat transfer in kilojoules will she generate in the process?

Exercise:

Problem:

Very large forces are produced in joints when a person jumps from some height to the ground. (a) Calculate the magnitude of the force produced if an 80.0-kg person jumps from a 0.600–m-high ledge and lands stiffly, compressing joint material 1.50 cm as a result. (Be certain to include the weight of the person.) (b) In practice the knees bend almost involuntarily to help extend the distance over which you stop. Calculate the magnitude of the force produced if the stopping distance is 0.300 m. (c) Compare both forces with the weight of the person.

Solution:

(a) 3.21×10^{4} N

(b) 2.35×10^{3} N

(c) Ratio of net force to weight of person is 41.0 in part (a); 3.00 in part (b)

Exercise:

Problem:

Jogging on hard surfaces with insufficiently padded shoes produces large forces in the feet and legs. (a) Calculate the magnitude of the force needed to stop the downward motion of a jogger’s leg, if his leg has a mass of 13.0 kg, a speed of 6.00 m/s, and stops in a distance of 1.50 cm. (Be certain to include the weight of the 75.0-kg jogger’s body.) (b) Compare this force with the weight of the jogger.

Exercise:

Problem:

(a) Calculate the energy in kJ used by a 55.0-kg woman who does 50 deep knee bends in which her center of mass is lowered and raised 0.400 m. (She does work in both directions.) You may assume her efficiency is 20%. (b) What is the average power consumption rate in watts if she does this in 3.00 min?

Solution:

(a) 108 kJ

(b) 599 W

Exercise:

Problem:

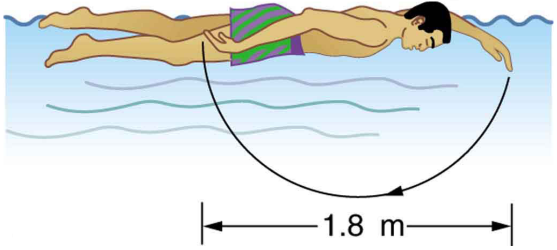
Kanellos Kanellopoulos flew 119 km from Crete to Santorini, Greece, on April 23, 1988, in the *Daedalus 88*, an aircraft powered by a bicycle-type drive mechanism (see [link]). His useful power output for the 234-min trip was about 350 W. Using the efficiency for cycling from [link], calculate the food energy in kilojoules he metabolized during the flight.

The Daedalus 88 in flight. (credit: NASA photo by Beasley)

Exercise:

Problem:

The swimmer shown in [link] exerts an average horizontal backward force of 80.0 N with his arm during each 1.80 m long stroke. (a) What is his work output in each stroke? (b) Calculate the power output of his arms if he does 120 strokes per minute.



Solution:

(a) 144 J

(b) 288 W

Exercise:

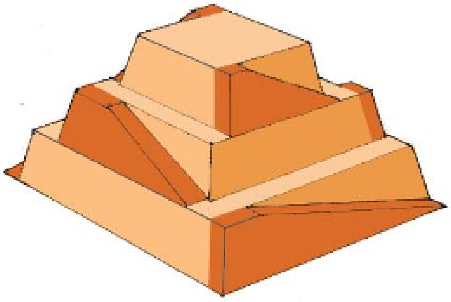
Problem:

Mountain climbers carry bottled oxygen when at very high altitudes. (a) Assuming that a mountain climber uses oxygen at twice the rate for climbing 116 stairs per minute (because of low air temperature and winds), calculate how many liters of oxygen a climber would need for 10.0 h of climbing. (These are liters at sea level.) Note that only 40% of the inhaled oxygen is utilized; the rest is exhaled. (b) How much useful work does the climber do if he and his equipment have a mass of 90.0 kg and he gains 1000 m of altitude? (c) What is his efficiency for the 10.0-h climb?

Exercise:

Problem:

The awe-inspiring Great Pyramid of Cheops was built more than 4500 years ago. Its square base, originally 230 m on a side, covered 13.1 acres, and it was 146 m high, with a mass of about 7×10^9 kg. (The pyramid’s dimensions are slightly different today due to quarrying and some sagging.) Historians estimate that 20,000 workers spent 20 years to construct it, working 12-hour days, 330 days per year. (a) Calculate the gravitational potential energy stored in the pyramid, given its center of mass is at one-fourth its height. (b) Only a fraction of the workers lifted blocks; most were involved in support services such as building ramps (see [link]), bringing food and water, and hauling blocks to the site. Calculate the efficiency of the workers who did the lifting, assuming there were 1000 of them and they consumed food energy at the rate of 300 kcal/h. What does your answer imply about how much of their work went into block-lifting, versus how much work went into friction and lifting and lowering their own bodies? (c) Calculate the mass of food that had to be supplied each day, assuming that the average worker required 3600 kcal per day and that their diet was 5% protein, 60% carbohydrate, and 35% fat. (These proportions neglect the mass of bulk and nondigestible materials consumed.)



Ancient pyramids were probably constructed using ramps as simple machines. (credit: Franck Monnier, Wikimedia Commons)

Solution:

(a) 2.50×10^{12} J

(b) 2.52%

(c) 1.4×10^{4} kg (14 metric tons)

Exercise:

Problem:

(a) How long can you play tennis on the 800 kJ (about 200 kcal) of energy in a candy bar? (b) Does this seem like a long time? Discuss why exercise is necessary but may not be sufficient to cause a person to lose weight.

Glossary

metabolic rate

the rate at which the body uses food energy to sustain life and to do different activities

basal metabolic rate

the total energy conversion rate of a person at rest

useful work

work done on an external system

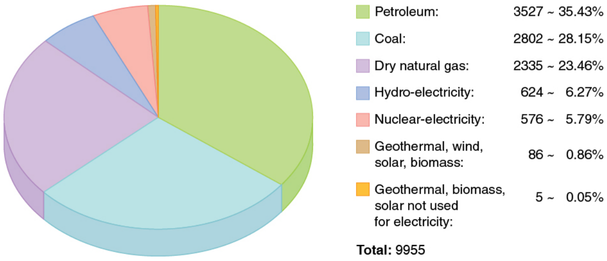
World Energy Use

- Describe the distinction between renewable and nonrenewable energy sources.
- Explain why the inevitable conversion of energy to less useful forms makes it necessary to conserve energy resources.

Energy is an important ingredient in all phases of society. We live in a very interdependent world, and access to adequate and reliable energy resources is crucial for economic growth and for maintaining the quality of our lives. But current levels of energy consumption and production are not sustainable. About 40% of the world’s energy comes from oil, and much of that goes to transportation uses. Oil prices are dependent as much upon new (or foreseen) discoveries as they are upon political events and situations around the world. The U.S., with 4.5% of the world’s population, consumes 24% of the world’s oil production per year; 66% of that oil is imported!

Renewable and Nonrenewable Energy Sources

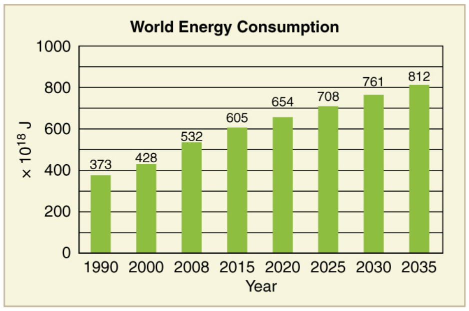
The principal energy resources used in the world are shown in [link]. The fuel mix has changed over the years but now is dominated by oil, although natural gas and solar contributions are increasing. **Renewable forms of energy** are those sources that cannot be used up, such as water, wind, solar, and biomass. About 85% of our energy comes from nonrenewable **fossil fuels**—oil, natural gas, coal. The likelihood of a link between global warming and fossil fuel use, with its production of carbon dioxide through combustion, has made, in the eyes of many scientists, a shift to non-fossil fuels of utmost importance—but it will not be easy.



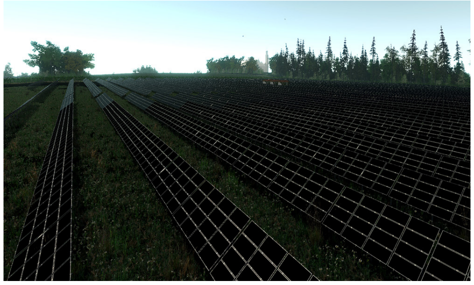
World energy consumption by source, in billions of kilowatt-hours: 2006. (credit: KVDP)

The World’s Growing Energy Needs

World energy consumption continues to rise, especially in the developing countries. (See [link].) Global demand for energy has tripled in the past 50 years and might triple again in the next 30 years. While much of this growth will come from the rapidly booming economies of China and India, many of the developed countries, especially those in Europe, are hoping to meet their energy needs by expanding the use of renewable sources. Although presently only a small percentage, renewable energy is growing very fast, especially wind energy. For example, Germany plans to meet 20% of its electricity and 10% of its overall energy needs with renewable resources by the year 2020. (See [link].) Energy is a key constraint in the rapid economic growth of China and India. In 2003, China surpassed Japan as the world’s second largest consumer of oil. However, over 1/3 of this is imported. Unlike most Western countries, coal dominates the commercial energy resources of China, accounting for 2/3 of its energy consumption. In 2009 China surpassed the United States as the largest generator of CO_2. In India, the main energy resources are biomass (wood and dung) and coal. Half of India’s oil is imported. About 70% of India’s electricity is generated by highly polluting coal. Yet there are sizeable strides being made in renewable energy. India has a rapidly growing wind energy base, and it has the largest solar cooking program in the world.



Past and projected world energy use (source: Based on data from U.S. Energy Information Administration, 2011)



Solar cell arrays at a power plant in Steindorf, Germany (credit: Michael Betke, Flickr)

[link] displays the 2006 commercial energy mix by country for some of the prime energy users in the world. While non-renewable sources dominate, some countries get a sizeable percentage of their electricity from renewable resources. For example, about 67% of New Zealand’s electricity demand is met by hydroelectric. Only 10% of the U.S. electricity is generated by renewable resources, primarily hydroelectric. It is difficult to determine total contributions of renewable energy in some countries with a large rural population, so these percentages in this table are left blank.

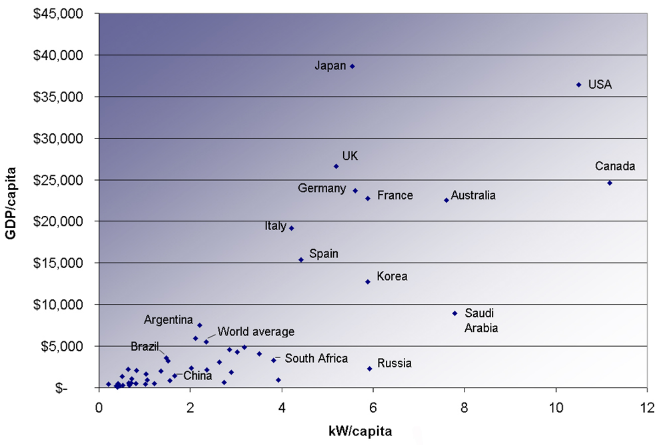
Country	Consumption, in EJ (10^{18} J)	Oil	Natural Gas	Coal	Nuclear	Hydro	Other Renewables
Australia	5.4	34%	17%	44%	0%	3%	1%

Country	Consumption, in EJ (10^{18} J)	Oil	Natural Gas	Coal	Nuclear	Hydro	Other Renewables
Brazil	9.6	48%	7%	5%	1%	35%	2%
China	63	22%	3%	69%	1%	6%	
Egypt	2.4	50%	41%	1%	0%	6%	
Germany	16	37%	24%	24%	11%	1%	3%
India	15	34%	7%	52%	1%	5%	
Indonesia	4.9	51%	26%	16%	0%	2%	3%
Japan	24	48%	14%	21%	12%	4%	1%
New Zealand	0.44	32%	26%	6%	0%	11%	19%
Russia	31	19%	53%	16%	5%	6%	
U.S.	105	40%	23%	22%	8%	3%	1%
World	**432**	**39%**	**23%**	**24%**	**6%**	**6%**	**2%**

Energy Consumption—Selected Countries (2006)

Energy and Economic Well-being

The last two columns in this table examine the energy and electricity use per capita. Economic well-being is dependent upon energy use, and in most countries higher standards of living, as measured by GDP (gross domestic product) per capita, are matched by higher levels of energy consumption per capita. This is borne out in [link]. Increased efficiency of energy use will change this dependency. A global problem is balancing energy resource development against the harmful effects upon the environment in its extraction and use.



Power consumption per capita versus GDP per capita for various countries. Note the increase in energy usage with increasing GDP. (2007, credit: Frank van Mierlo, Wikimedia Commons)

Conserving Energy

As we finish this chapter on energy and work, it is relevant to draw some distinctions between two sometimes misunderstood terms in the area of energy use. As has been mentioned elsewhere, the “law of the conservation of energy” is a very useful principle in analyzing physical processes. It is a statement that cannot be proven from basic principles, but is a very good bookkeeping device, and no exceptions have ever been found. It states that the total amount of energy in an isolated system will always remain constant. Related to this principle, but remarkably different from it, is the important philosophy of energy conservation. This concept has to do with seeking to decrease the amount of energy used by an individual or group through (1) reduced activities (e.g., turning down thermostats, driving fewer kilometers) and/or (2) increasing conversion efficiencies in the performance of a particular task—such as developing and using more efficient room heaters, cars that have greater miles-per-gallon ratings, energy-efficient compact fluorescent lights, etc.

Since energy in an isolated system is not destroyed or created or generated, one might wonder why we need to be concerned about our energy resources, since energy is a conserved quantity. The problem is that the final result of most energy transformations is waste heat transfer to the environment and conversion to energy forms no longer useful for doing work. To state it in another way, the potential for energy to produce useful work has been “degraded” in the energy transformation. (This will be discussed in more detail in Thermodynamics.)

Section Summary

- The relative use of different fuels to provide energy has changed over the years, but fuel use is currently dominated by oil, although natural gas and solar contributions are increasing.
- Although non-renewable sources dominate, some countries meet a sizeable percentage of their electricity needs from renewable resources.
- The United States obtains only about 10% of its energy from renewable sources, mostly hydroelectric power.
- Economic well-being is dependent upon energy use, and in most countries higher standards of living, as measured by GDP (Gross Domestic Product) per capita, are matched by higher levels of energy consumption per capita.
- Even though, in accordance with the law of conservation of energy, energy can never be created or destroyed, energy that can be used to do work is always partly converted to less useful forms, such as waste heat to the environment, in all of our uses of energy for practical purposes.

Conceptual Questions

Exercise:

Problem:

What is the difference between energy conservation and the law of conservation of energy? Give some examples of each.

Exercise:

Problem:

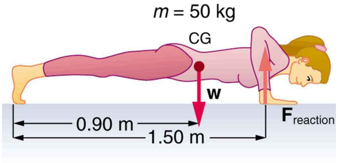
If the efficiency of a coal-fired electrical generating plant is 35%, then what do we mean when we say that energy is a conserved quantity?

Problems & Exercises

Exercise:

Problem: Integrated Concepts

(a) Calculate the force the woman in [link] exerts to do a push-up at constant speed, taking all data to be known to three digits. (b) How much work does she do if her center of mass rises 0.240 m? (c) What is her useful power output if she does 25 push-ups in 1 min? (Should work done lowering her body be included? See the discussion of useful work in Work, Energy, and Power in Humans.



Forces involved in doing push-ups. The woman’s weight acts as a force exerted downward on her center of gravity (CG).

Solution:

(a) 294 N

(b) 118 J

(c) 49.0 W

Exercise:

Problem: Integrated Concepts

A 75.0-kg cross-country skier is climbing a 3.0° slope at a constant speed of 2.00 m/s and encounters air resistance of 25.0 N. Find his power output for work done against the gravitational force and air resistance. (b) What average force does he exert backward on the snow to accomplish this? (c) If he continues to exert this force and to experience the same air resistance when he reaches a level area, how long will it take him to reach a velocity of 10.0 m/s?

Exercise:

Problem: **Integrated Concepts**

The 70.0-kg swimmer in [link] starts a race with an initial velocity of 1.25 m/s and exerts an average force of 80.0 N backward with his arms during each 1.80 m long stroke. (a) What is his initial acceleration if water resistance is 45.0 N? (b) What is the subsequent average resistance force from the water during the 5.00 s it takes him to reach his top velocity of 2.50 m/s? (c) Discuss whether water resistance seems to increase linearly with velocity.

Solution:

(a) $0.500\ \mathrm{m/s}^2$

(b) $62.5\ \mathrm{N}$

(c) Assuming the acceleration of the swimmer decreases linearly with time over the 5.00 s interval, the frictional force must therefore be increasing linearly with time, since $f = F - \mathrm{ma}$. If the acceleration decreases linearly with time, the velocity will contain a term dependent on time squared (t^2). Therefore, the water resistance will not depend linearly on the velocity.

Exercise:

Problem: **Integrated Concepts**

A toy gun uses a spring with a force constant of 300 N/m to propel a 10.0-g steel ball. If the spring is compressed 7.00 cm and friction is negligible: (a) How much force is needed to compress the spring? (b) To what maximum height can the ball be shot? (c) At what angles above the horizontal may a child aim to hit a target 3.00 m away at the same height as the gun? (d) What is the gun’s maximum range on level ground?

Exercise:

Problem: **Integrated Concepts**

(a) What force must be supplied by an elevator cable to produce an acceleration of $0.800\ \mathrm{m/s}^2$ against a 200-N frictional force, if the mass of the loaded elevator is 1500 kg? (b) How much work is done by the cable in lifting the elevator 20.0 m? (c) What is the final speed of the elevator if it starts from rest? (d) How much work went into thermal energy?

Solution:

(a) $16.1 \times 10^3\ \mathrm{N}$

(b) $3.22 \times 10^5\ \mathrm{J}$

(c) 5.66 m/s

(d) 4.00 kJ

Exercise:

Problem: **Unreasonable Results**

A car advertisement claims that its 900-kg car accelerated from rest to 30.0 m/s and drove 100 km, gaining 3.00 km in altitude, on 1.0 gal of gasoline. The average force of friction including air resistance was 700 N. Assume all values are known to three significant figures. (a) Calculate the car’s efficiency. (b) What is unreasonable about the result? (c) Which premise is unreasonable, or which premises are inconsistent?

Exercise:

Problem: Unreasonable Results

Body fat is metabolized, supplying 9.30 kcal/g, when dietary intake is less than needed to fuel metabolism. The manufacturers of an exercise bicycle claim that you can lose 0.500 kg of fat per day by vigorously exercising for 2.00 h per day on their machine. (a) How many kcal are supplied by the metabolization of 0.500 kg of fat? (b) Calculate the kcal/min that you would have to utilize to metabolize fat at the rate of 0.500 kg in 2.00 h. (c) What is unreasonable about the results? (d) Which premise is unreasonable, or which premises are inconsistent?

Solution:

(a) 4.65×10^{3} kcal

(b) 38.8 kcal/min

(c) This power output is higher than the highest value on [link], which is about 35 kcal/min (corresponding to 2415 watts) for sprinting.

(d) It would be impossible to maintain this power output for 2 hours (imagine sprinting for 2 hours!).

Exercise:

Problem: Construct Your Own Problem

Consider a person climbing and descending stairs. Construct a problem in which you calculate the long-term rate at which stairs can be climbed considering the mass of the person, his ability to generate power with his legs, and the height of a single stair step. Also consider why the same person can descend stairs at a faster rate for a nearly unlimited time in spite of the fact that very similar forces are exerted going down as going up. (This points to a fundamentally different process for descending versus climbing stairs.)

Exercise:

Problem: Construct Your Own Problem

Consider humans generating electricity by pedaling a device similar to a stationary bicycle. Construct a problem in which you determine the number of people it would take to replace a large electrical generation facility. Among the things to consider are the power output that is reasonable using the legs, rest time, and the need for electricity 24 hours per day. Discuss the practical implications of your results.

Exercise:

Problem: Integrated Concepts

A 105-kg basketball player crouches down 0.400 m while waiting to jump. After exerting a force on the floor through this 0.400 m, his feet leave the floor and his center of gravity rises 0.950 m above its normal standing erect position. (a) Using energy considerations, calculate his velocity when he leaves the floor. (b) What average force did he exert on the floor? (Do not neglect the force to support his weight as well as that to accelerate him.) (c) What was his power output during the acceleration phase?

Solution:

(a) 4.32 m/s

(b) 3.47×10^{3} N

(c) 8.93 kW

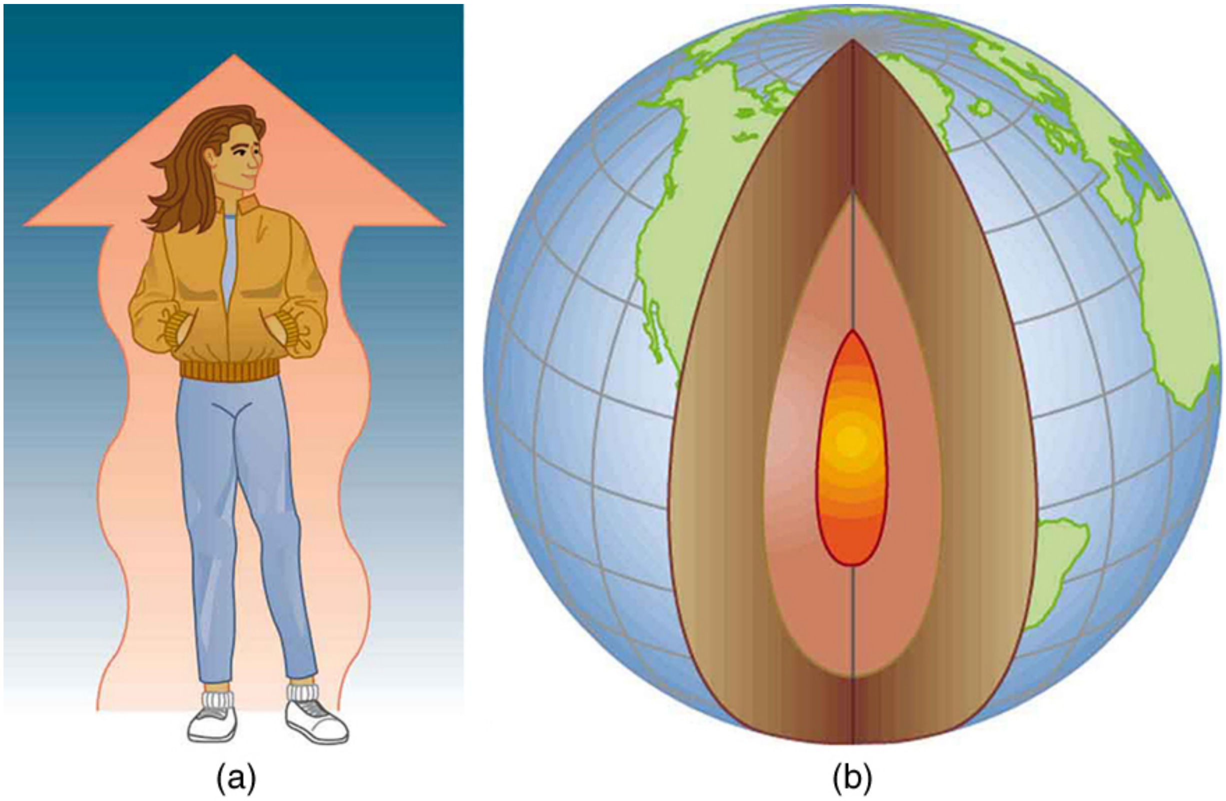
Glossary

renewable forms of energy
: those sources that cannot be used up, such as water, wind, solar, and biomass

fossil fuels
: oil, natural gas, and coal

Introduction to Heat and Heat Transfer Methods class="introduction"

(a) The chilling effect of a clear breezy night is produced by the wind and by radiative heat transfer to cold outer space. (b) There was once great controversy about the Earth’s age, but it is now generally accepted to be about 4.5 billion years old. Much of the debate is centered on the Earth’s molten interior. According to our understanding of heat transfer, if the Earth is really that old, its center should have cooled off long ago. The discovery of radioactivity in rocks revealed the source of energy that keeps the Earth’s interior molten, despite heat transfer to the surface, and from there to cold outer space.



Energy can exist in many forms and heat is one of the most intriguing. Heat is often hidden, as it only exists when in transit, and is transferred by a number of distinctly different methods. Heat transfer touches every aspect of our lives and helps us understand how the universe functions. It explains the chill we feel on a clear breezy night, or why Earth’s core has yet to cool. This chapter defines and explores heat transfer, its effects, and the methods by which heat is transferred. These topics are fundamental, as well as practical, and will often be referred to in the chapters ahead.

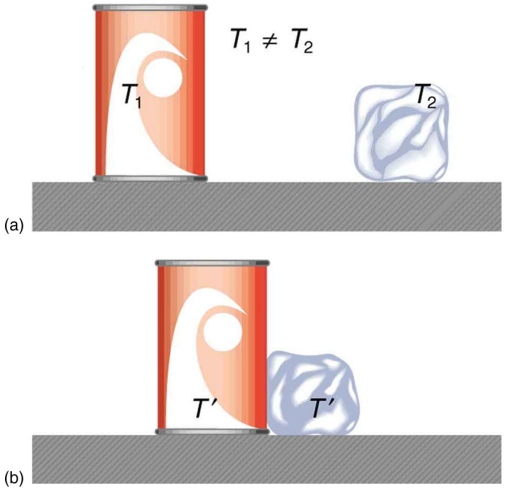
Heat

- Define heat as transfer of energy.

In Work, Energy, and Energy Resources, we defined work as force times distance and learned that work done on an object changes its kinetic energy. We also saw in Temperature, Kinetic Theory, and the Gas Laws that temperature is proportional to the (average) kinetic energy of atoms and molecules. We say that a thermal system has a certain internal energy: its internal energy is higher if the temperature is higher. If two objects at different temperatures are brought in contact with each other, energy is transferred from the hotter to the colder object until equilibrium is reached and the bodies reach thermal equilibrium (i.e., they are at the same temperature). No work is done by either object, because no force acts through a distance. The transfer of energy is caused by the temperature difference, and ceases once the temperatures are equal. These observations lead to the following definition of **heat**: Heat is the spontaneous transfer of energy due to a temperature difference.

As noted in Temperature, Kinetic Theory, and the Gas Laws, heat is often confused with temperature. For example, we may say the heat was unbearable, when we actually mean that the temperature was high. Heat is a form of energy, whereas temperature is not. The misconception arises because we are sensitive to the flow of heat, rather than the temperature.

Owing to the fact that heat is a form of energy, it has the SI unit of *joule* (J). The *calorie* (cal) is a common unit of energy, defined as the energy needed to change the temperature of 1.00 g of water by 1.00^{o}C —specifically, between 14.5^{o}C and 15.5^{o}C, since there is a slight temperature dependence. Perhaps the most common unit of heat is the **kilocalorie** (kcal), which is the energy needed to change the temperature of 1.00 kg of water by 1.00^{o}C. Since mass is most often specified in kilograms, kilocalorie is commonly used. Food calories (given the notation Cal, and sometimes called "big calorie") are actually kilocalories ($1\text{ kilocalorie}=1000\text{ calories}$), a fact not easily determined from package labeling.



In figure (a) the soft drink and the ice have different temperatures, T_1 and T_2, and are not in thermal equilibrium. In figure (b), when the soft drink and ice are allowed to interact, energy is transferred until they reach the same temperature $T\prime$, achieving equilibrium. Heat transfer occurs due to the difference in temperatures. In fact, since the soft drink and ice are both in contact with the surrounding air and bench, the equilibrium temperature will be the same for both.

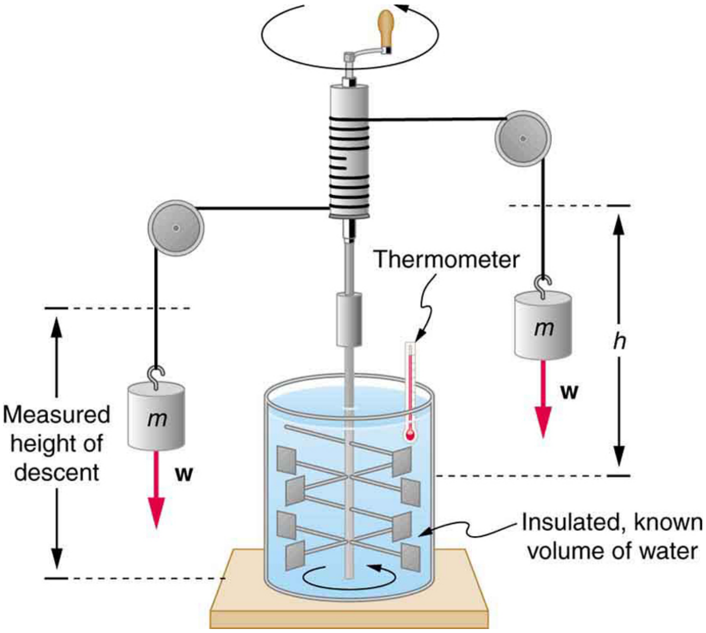
Mechanical Equivalent of Heat

It is also possible to change the temperature of a substance by doing work. Work can transfer energy into or out of a system. This realization helped establish the fact that heat is a form of energy. James Prescott Joule (1818–1889) performed many experiments to establish the **mechanical equivalent of heat**—*the work needed to produce the same effects as heat transfer*. In terms of the units used for these two terms, the best modern value for this equivalence is

Equation:

$$1.000\text{ kcal} = 4186\text{ J}.$$

We consider this equation as the conversion between two different units of energy.



Schematic depiction of Joule’s experiment that established the equivalence of heat and work.

The figure above shows one of Joule's most famous experimental setups for demonstrating the mechanical equivalent of heat. It demonstrated that work and heat can produce the same effects, and helped establish the principle of conservation of energy. Gravitational potential energy (PE) (work done by the gravitational force) is converted into kinetic energy (KE), and then randomized by viscosity and turbulence into increased average kinetic energy of atoms and molecules in the system, producing a temperature increase. His contributions to the field of thermodynamics were so significant that the SI unit of energy was named after him.

Heat added or removed from a system changes its internal energy and thus its temperature. Such a temperature increase is observed while cooking. However, adding heat does not necessarily increase the temperature. An example is melting of ice; that is, when a substance changes from one phase to another. Work done on the system or by the system can also change the internal energy of the system. Joule demonstrated that the temperature of a system can be increased by stirring. If an ice cube is rubbed against a rough surface, work is done by the frictional force. A system has a well-defined internal energy, but we cannot say that it has a certain "heat content" or "work content". We use the phrase "heat transfer" to emphasize its nature.

Exercise:

Check Your Understanding

Problem:

Two samples (A and B) of the same substance are kept in a lab. Someone adds 10 kilojoules (kJ) of heat to one sample, while 10 kJ of work is done on the other sample. How can you tell to which sample the heat was added?

Solution:

Heat and work both change the internal energy of the substance. However, the properties of the sample only depend on the internal energy so that it is impossible to tell whether heat was added to sample A or B.

Summary

- Heat and work are the two distinct methods of energy transfer.
- Heat is energy transferred solely due to a temperature difference.
- Any energy unit can be used for heat transfer, and the most common are kilocalorie (kcal) and joule (J).
- Kilocalorie is defined to be the energy needed to change the temperature of 1.00 kg of water between 14.5°C and 15.5°C.
- The mechanical equivalent of this heat transfer is $1.00 \text{ kcal } = \text{ } 4186 \text{ J}$.

Conceptual Questions

Exercise:

Problem: How is heat transfer related to temperature?

Exercise:

Problem:

Describe a situation in which heat transfer occurs. What are the resulting forms of energy?

Exercise:

Problem:

When heat transfers into a system, is the energy stored as heat? Explain briefly.

Glossary

heat
: the spontaneous transfer of energy due to a temperature difference

kilocalorie
: $1 \text{ kilocalorie} = 1000 \text{ calories}$

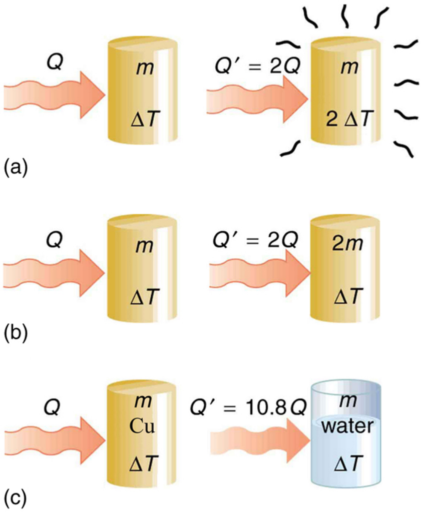
mechanical equivalent of heat

the work needed to produce the same effects as heat transfer

Temperature Change and Heat Capacity

- Observe heat transfer and change in temperature and mass.
- Calculate final temperature after heat transfer between two objects.

One of the major effects of heat transfer is temperature change: heating increases the temperature while cooling decreases it. We assume that there is no phase change and that no work is done on or by the system. Experiments show that the transferred heat depends on three factors—the change in temperature, the mass of the system, and the substance and phase of the substance.



The heat Q transferred to cause a temperature change depends on the magnitude of the temperature change, the mass of the system, and the substance and phase involved. (a) The amount of heat transferred is directly proportional to the temperature change. To double the temperature change of a mass m, you need to add twice the heat. (b) The amount of heat transferred is also directly proportional to the mass. To cause an equivalent temperature change in a doubled mass, you need to add twice the heat. (c) The amount of heat transferred depends on the substance and its phase. If it takes an amount Q of heat to cause a temperature change ΔT in a given mass of copper, it will take 10.8 times that amount of heat to cause the equivalent temperature change in the same mass of water assuming no phase change in either substance.

The dependence on temperature change and mass are easily understood. Owing to the fact that the (average) kinetic energy of an atom or molecule is proportional to the absolute temperature, the internal energy of a system is proportional to the absolute temperature and the number of atoms or molecules. Owing to the fact that the transferred heat is equal to the change in the internal energy, the heat is proportional to the mass of the substance and the temperature change. The transferred heat also depends on the substance so that, for example, the heat necessary to raise the temperature is less for alcohol than for water. For the same substance, the transferred heat also depends on the phase (gas, liquid, or solid).

Note:
Heat Transfer and Temperature Change
The quantitative relationship between heat transfer and temperature change contains all three factors:
Equation:

$$Q=\text{mc}\Delta T,$$

where Q is the symbol for heat transfer, m is the mass of the substance, and ΔT is the change in temperature. The symbol c stands for **specific heat** and depends on the material and phase. The specific heat is the amount of heat necessary to change the temperature of 1.00 kg of mass by 1.00°C. The specific heat c is a property of the substance; its SI unit is $\text{J}/(\text{kg}\cdot\text{K})$ or $\text{J}/(\text{kg}\cdot{}^{\circ}\text{C})$. Recall that the temperature change (ΔT) is the same in units of kelvin and degrees Celsius. If heat transfer is measured in kilocalories, then *the unit of specific heat* is $\text{kcal}/(\text{kg}\cdot{}^{\circ}\text{C})$.

Values of specific heat must generally be looked up in tables, because there is no simple way to calculate them. In general, the specific heat also depends on the temperature. [link] lists representative values of specific heat for various substances. Except for gases, the temperature and volume dependence of the specific heat of most substances is weak. We see from this table that the specific heat of water is five times that of glass and ten times that of iron, which means that it takes five times as much heat to raise the temperature of water the same amount as for glass and ten times as much heat to raise the temperature of water as for iron. In fact, water has one of the largest specific heats of any material, which is important for sustaining life on Earth.

Example:
Calculating the Required Heat: Heating Water in an Aluminum Pan

A 0.500 kg aluminum pan on a stove is used to heat 0.250 liters of water from 20.0°C to 80.0°C. (a) How much heat is required? What percentage of the heat is used to raise the temperature of (b) the pan and (c) the water?

Strategy

The pan and the water are always at the same temperature. When you put the pan on the stove, the temperature of the water and the pan is increased by the same amount. We use the equation for the heat transfer for the given temperature change and mass of water and aluminum. The specific heat values for water and aluminum are given in [link].

Solution

Because water is in thermal contact with the aluminum, the pan and the water are at the same temperature.

1. Calculate the temperature difference:
   Equation:

$$\Delta T = T_{\text{f}} - T_{\text{i}} = 60.0^{\circ}\text{C}.$$

2. Calculate the mass of water. Because the density of water is $1000\ \text{kg}/\text{m}^3$, one liter of water has a mass of 1 kg, and the mass of 0.250 liters of water is $m_{\text{w}} = 0.250\ \text{kg}$.
3. Calculate the heat transferred to the water. Use the specific heat of water in [link]:
   Equation:

$$Q_{\mathrm{w}} = m_{\mathrm{w}}c_{\mathrm{w}}\Delta T = (0.250\text{ kg})(4186\text{ J/kg}^{\mathrm{o}}\mathrm{C})(60.0^{\mathrm{o}}\mathrm{C}) = 62.8\text{ kJ}.$$

4. Calculate the heat transferred to the aluminum. Use the specific heat for aluminum in [link]:
   Equation:

   $$Q_{\mathrm{Al}} = m_{\mathrm{Al}}c_{\mathrm{Al}}\Delta T = (0.500\text{ kg})(900\text{ J/kg}^{\mathrm{o}}\mathrm{C})(60.0^{\mathrm{o}}\mathrm{C}) = 27.0 \times 10^{4}\mathrm{J} = 27.0\text{ kJ}.$$

5. Compare the percentage of heat going into the pan versus that going into the water. First, find the total transferred heat:
   Equation:

   $$Q_{\text{Total}} = Q_{\mathrm{W}} + Q_{\mathrm{Al}} = 62.8\text{ kJ} + 27.0\ \text{ kJ} = 89.8\text{ kJ}.$$

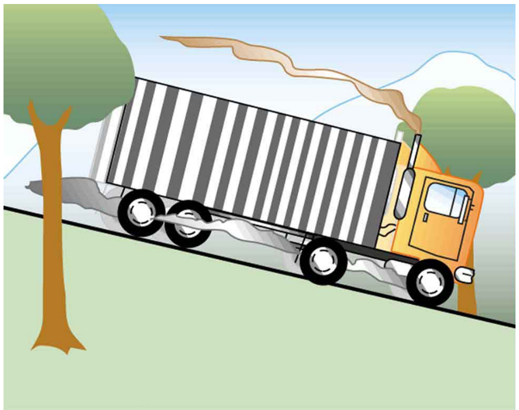
Thus, the amount of heat going into heating the pan is
Equation:

$$\frac{27.0\text{ kJ}}{89.8\text{ kJ}} \times 100\% = 30.1\%,$$

and the amount going into heating the water is
Equation:

$$\frac{62.8\text{ kJ}}{89.8\text{ kJ}} \times 100\% = 69.9\%.$$

Discussion
In this example, the heat transferred to the container is a significant fraction of the total transferred heat. Although the mass of the pan is twice that of the water, the specific heat of water is over four times greater than that of aluminum. Therefore, it takes a bit more than twice the heat to achieve the given temperature change for the water as compared to the aluminum pan.



The smoking brakes on this truck are a visible evidence of the mechanical equivalent of heat.

Example:

Calculating the Temperature Increase from the Work Done on a Substance: Truck Brakes Overheat on Downhill Runs

Truck brakes used to control speed on a downhill run do work, converting gravitational potential energy into increased internal energy (higher temperature) of the brake material. This conversion prevents the gravitational potential energy from being converted into kinetic energy of the truck. The problem is that the mass of the truck is large compared with that of the brake material absorbing the energy, and the temperature increase may occur too fast for sufficient heat to transfer from the brakes to the environment.

Calculate the temperature increase of 100 kg of brake material with an average specific heat of 800 $\text{J/kg}\cdot{}^{\text{o}}\text{C}$ if the material retains 10% of the energy from a 10,000-kg truck descending 75.0 m (in vertical displacement) at a constant speed.

Strategy

If the brakes are not applied, gravitational potential energy is converted into kinetic energy. When brakes are applied, gravitational potential energy is converted into internal energy of the brake material. We first calculate the gravitational potential energy (Mgh) that the entire truck loses in its descent and then find the temperature increase produced in the brake material alone.

Solution

1. Calculate the change in gravitational potential energy as the truck goes downhill
   Equation:

$$\text{Mgh} = (10{,}000\ \text{kg})\left(9.80\ \text{m/s}^2\right)(75.0\ \text{m}) = 7.35 \times 10^6\ \text{J}.$$

2. Calculate the temperature from the heat transferred using Q=Mgh and
   Equation:

$$\Delta T = \frac{Q}{\text{mc}},$$

where m is the mass of the brake material. Insert the values $m = 100\ \text{kg}$ and $c = 800\ \text{J/kg} \cdot {}^\text{o}\text{C}$ to find
Equation:

$$\Delta T = \frac{\left(7.35 \times 10^5\ \text{J}\right)}{(100\ \text{kg})(800\ \text{J/kg}^\text{o}\text{C})} = 9.2^\text{o}\text{C}.$$

Discussion
This same idea underlies the recent hybrid technology of cars, where mechanical energy (gravitational potential energy) is converted by the brakes into electrical energy (battery).

Substances	**Specific heat (*c*)**	
Solids	J/kg·°C	kcal/kg·°C[footnote] These values are identical in units of cal/g ·°C.
Aluminum	900	0.215
Asbestos	800	0.19
Concrete, granite (average)	840	0.20
Copper	387	0.0924
Glass	840	0.20

Substances	**Specific heat (*c*)**	
Gold	129	0.0308
Human body (average at 37 °C)	3500	0.83
Ice (average, -50°C to 0°C)	2090	0.50
Iron, steel	452	0.108
Lead	128	0.0305
Silver	235	0.0562
Wood	1700	0.4
Liquids		
Benzene	1740	0.415
Ethanol	2450	0.586
Glycerin	2410	0.576
Mercury	139	0.0333
Water (15.0 °C)	4186	1.000
Gases [footnote] c_v at constant volume and at 20.0°C, except as noted, and at 1.00 atm average pressure. Values in parentheses are c_p at a constant pressure of 1.00 atm.		
Air (dry)	721 (1015)	0.172 (0.242)
Ammonia	1670 (2190)	0.399 (0.523)
Carbon dioxide	638 (833)	0.152 (0.199)

Substances	**Specific heat (*c*)**	
Nitrogen	739 (1040)	0.177 (0.248)
Oxygen	651 (913)	0.156 (0.218)
Steam (100°C)	1520 (2020)	0.363 (0.482)

Specific Heats[footnote] of Various Substances
The values for solids and liquids are at constant volume and at $25^{\circ}\mathrm{C}$, except as noted.

Note that [link] is an illustration of the mechanical equivalent of heat. Alternatively, the temperature increase could be produced by a blow torch instead of mechanically.

Example:
Calculating the Final Temperature When Heat Is Transferred Between Two Bodies: Pouring Cold Water in a Hot Pan
Suppose you pour 0.250 kg of $20.0^{\circ}\mathrm{C}$ water (about a cup) into a 0.500-kg aluminum pan off the stove with a temperature of $150^{\circ}\mathrm{C}$. Assume that the pan is placed on an insulated pad and that a negligible amount of water boils off. What is the temperature when the water and pan reach thermal equilibrium a short time later?
Strategy
The pan is placed on an insulated pad so that little heat transfer occurs with the surroundings. Originally the pan and water are not in thermal equilibrium: the pan is at a higher temperature than the water. Heat transfer then restores thermal equilibrium once the water and pan are in contact. Because heat transfer between the pan and water takes place rapidly, the mass of evaporated water is negligible and the magnitude of the heat lost by the pan is equal to the heat gained by the water. The exchange of heat stops once a thermal equilibrium between the pan and the water is achieved. The heat exchange can be written as $\mid Q_{\text{hot}} \mid = Q_{\text{cold}}$.
Solution

1. Use the equation for heat transfer $Q = \mathrm{mc}\Delta T$ to express the heat lost by the aluminum pan in terms of the mass of the pan, the specific heat of aluminum, the initial temperature of the pan, and the final temperature:
   Equation:

$$Q_{\text{hot}} = m_{\text{Al}} c_{\text{Al}} (T_{\text{f}} - 150^{\circ}\mathrm{C}).$$

2. Express the heat gained by the water in terms of the mass of the water, the specific heat of water, the initial temperature of the water and the final temperature:
   Equation:

$$Q_{\text{cold}} = m_{\text{W}}c_{\text{W}}(T_{\text{f}} - 20.0^{\text{o}}\text{C}).$$

3. Note that $Q_{\text{hot}} < 0$ and $Q_{\text{cold}} > 0$ and that they must sum to zero because the heat lost by the hot pan must be the same as the heat gained by the cold water:
   Equation:

$$\begin{aligned} Q_{\text{cold}} + Q_{\text{hot}} &= 0, \\ Q_{\text{cold}} &= -\text{Q}_{\text{hot}}, \\ m_{\text{W}}c_{\text{W}}(T_{\text{f}} - 20.0^{\text{o}}\text{C}) &= -\text{m}_{\text{Al}}c_{\text{Al}}(T_{\text{f}} - 150^{\text{o}}\text{C}.) \end{aligned}$$

4. This an equation for the unknown final temperature, T_{f}
5. Bring all terms involving T_{f} on the left hand side and all other terms on the right hand side. Solve for T_{f},
   Equation:

$$T_{\text{f}} = \frac{m_{\text{Al}}c_{\text{Al}}(150^{\text{o}}\text{C}) + m_{\text{W}}c_{\text{W}}(20.0^{\text{o}}\text{C})}{m_{\text{Al}}c_{\text{Al}} + m_{\text{W}}c_{\text{W}}},$$

   and insert the numerical values:
   Equation:

$$\begin{aligned} T_{\text{f}} &= \frac{(0.500\text{ kg})(900\text{ J/kg}^{\text{o}}\text{C})(150^{\text{o}}\text{C}) + (0.250\text{ kg})(4186\text{ J/kg}^{\text{o}}\text{C})(20.0^{\text{o}}\text{C})}{(0.500\text{ kg})(900\text{ J/kg}^{\text{o}}\text{C}) + (0.250\text{ kg})(4186\text{ J/kg}^{\text{o}}\text{C})} \\ &= \frac{88430\text{ J}}{1496.5\text{ J/}^{\text{o}}\text{C}} \\ &= 59.1^{\text{o}}\text{C}. \end{aligned}$$

Discussion

This is a typical *calorimetry* problem—two bodies at different temperatures are brought in contact with each other and exchange heat until a common temperature is reached. Why is the final temperature so much closer to 20.0^{o}C than 150^{o}C? The reason is that water has a greater specific heat than most common substances and thus undergoes a small temperature change for a given heat transfer. A large body of water, such as a lake, requires a large amount of heat to increase its temperature appreciably. This explains why the temperature of a lake stays relatively constant during a day even when the temperature change of the air is large. However, the water temperature does change over longer times (e.g., summer to winter).

Note:
Take-Home Experiment: Temperature Change of Land and Water
What heats faster, land or water?
To study differences in heat capacity:

- Place equal masses of dry sand (or soil) and water at the same temperature into two small jars. (The average density of soil or sand is about 1.6 times that of water, so you can achieve approximately equal masses by using 50% more water by volume.)
- Heat both (using an oven or a heat lamp) for the same amount of time.
- Record the final temperature of the two masses.
- Now bring both jars to the same temperature by heating for a longer period of time.
- Remove the jars from the heat source and measure their temperature every 5 minutes for about 30 minutes.

Which sample cools off the fastest? This activity replicates the phenomena responsible for land breezes and sea breezes.

Exercise:
Check Your Understanding

Problem:

If 25 kJ is necessary to raise the temperature of a block from 25ºC to 30ºC, how much heat is necessary to heat the block from 45ºC to 50ºC?

Solution:

The heat transfer depends only on the temperature difference. Since the temperature differences are the same in both cases, the same 25 kJ is necessary in the second case.

Summary

- The transfer of heat Q that leads to a change ΔT in the temperature of a body with mass m is $Q = \text{mc}\Delta T$, where c is the specific heat of the material. This relationship can also be considered as the definition of specific heat.

Conceptual Questions

Exercise:

Problem:

What three factors affect the heat transfer that is necessary to change an object’s temperature?

Exercise:

Problem:

The brakes in a car increase in temperature by ΔT when bringing the car to rest from a speed v. How much greater would ΔT be if the car initially had twice the speed? You may assume the car to stop sufficiently fast so that no heat transfers out of the brakes.

Problems & Exercises

Exercise:

Problem:

On a hot day, the temperature of an 80,000-L swimming pool increases by 1.50^{o}C. What is the net heat transfer during this heating? Ignore any complications, such as loss of water by evaporation.

Solution:
Equation:

$$5.02\times 10^{8}\text{ J}$$

Exercise:

Problem: Show that $1\text{ cal/g}\cdot {}^{\text{o}}\text{C}=1\text{ kcal/kg}\cdot {}^{\text{o}}\text{C}$.

Exercise:

Problem:

To sterilize a 50.0-g glass baby bottle, we must raise its temperature from 22.0^{o}C to 95.0^{o}C. How much heat transfer is required?

Solution:
Equation:

$$3.07\times 10^{3}\text{ J}$$

Exercise:

Problem:

The same heat transfer into identical masses of different substances produces different temperature changes. Calculate the final temperature when 1.00 kcal of heat transfers into 1.00 kg of the following, originally at $20.0^{\circ}C$: (a) water; (b) concrete; (c) steel; and (d) mercury.

Exercise:

Problem:

Rubbing your hands together warms them by converting work into thermal energy. If a woman rubs her hands back and forth for a total of 20 rubs, at a distance of 7.50 cm per rub, and with an average frictional force of 40.0 N, what is the temperature increase? The mass of tissues warmed is only 0.100 kg, mostly in the palms and fingers.

Solution:
Equation:

$$0.171^{\circ}C$$

Exercise:

Problem:

A 0.250-kg block of a pure material is heated from $20.0^{\circ}C$ to $65.0^{\circ}C$ by the addition of 4.35 kJ of energy. Calculate its specific heat and identify the substance of which it is most likely composed.

Exercise:

Problem:

Suppose identical amounts of heat transfer into different masses of copper and water, causing identical changes in temperature. What is the ratio of the mass of copper to water?

Solution:

10.8

Exercise:

Problem:

(a) The number of kilocalories in food is determined by calorimetry techniques in which the food is burned and the amount of heat transfer is measured. How many kilocalories per gram are there in a 5.00-g peanut if the energy from burning it is transferred to 0.500 kg of water held in a 0.100-kg aluminum cup, causing a 54.9^{o}C temperature increase? (b) Compare your answer to labeling information found on a package of peanuts and comment on whether the values are consistent.

Exercise:

Problem:

Following vigorous exercise, the body temperature of an 80.0-kg person is 40.0^{o}C. At what rate in watts must the person transfer thermal energy to reduce the the body temperature to 37.0^{o}C in 30.0 min, assuming the body continues to produce energy at the rate of 150 W? $(1 \text{ watt} = 1 \text{ joule/second or } 1 \text{ W} = 1 \text{ J/s })$.

Solution:

617 W

Exercise:

Problem:

Even when shut down after a period of normal use, a large commercial nuclear reactor transfers thermal energy at the rate of 150 MW by the radioactive decay of fission products. This heat transfer causes a rapid increase in temperature if the cooling system fails $(1 \text{ watt} = 1 \text{ joule/second or } 1 \text{ W} = 1 \text{ J/s and } 1 \text{ MW} = 1 \text{ megawatt})$. (a) Calculate the rate of temperature increase in degrees Celsius per second $(^{\text{o}}\text{C/s})$ if the mass of the reactor core is $1.60 \times 10^{5} \text{ kg}$ and it has an average specific heat of $0.3349 \text{ kJ/kg}^{\text{o}} \cdot \text{C}$. (b) How long would it take to obtain a temperature increase of 2000^{o}C, which could cause some metals holding the radioactive materials to melt? (The initial rate of temperature increase would be greater than that calculated here because the heat transfer is concentrated in a smaller mass. Later, however, the temperature increase would slow down because the 5×10^{5}-kg steel containment vessel would also begin to heat up.)

Radioactive spent-fuel pool at a nuclear power plant. Spent fuel stays hot for a long time. (credit: U.S. Department of Energy)

Glossary

specific heat

the amount of heat necessary to change the temperature of 1.00 kg of a substance by 1.00 ºC

Phase Change and Latent Heat

- Examine heat transfer.
- Calculate final temperature from heat transfer.

So far we have discussed temperature change due to heat transfer. No temperature change occurs from heat transfer if ice melts and becomes liquid water (i.e., during a phase change). For example, consider water dripping from icicles melting on a roof warmed by the Sun. Conversely, water freezes in an ice tray cooled by lower-temperature surroundings.

Heat from the air transfers to the ice causing it to melt. (credit: Mike Brand)

Energy is required to melt a solid because the cohesive bonds between the molecules in the solid must be broken apart such that, in the liquid, the molecules can move around at comparable kinetic energies; thus, there is no rise in temperature. Similarly, energy is needed to vaporize a liquid, because molecules in a liquid interact with each other via attractive forces. There is no temperature change until a phase change is complete. The temperature of a cup of soda initially at 0ºC stays at 0ºC until all the ice has melted. Conversely, energy is released during freezing and condensation, usually in the form of thermal energy. Work is done by cohesive forces when molecules are brought together. The corresponding energy must be given off (dissipated) to allow them to stay together [link].

The energy involved in a phase change depends on two major factors: the number and strength of bonds or force pairs. The number of bonds is proportional to the number of molecules and thus to the mass of the sample. The strength of forces depends on the type of molecules. The heat Q required to change the phase of a sample of mass m is given by

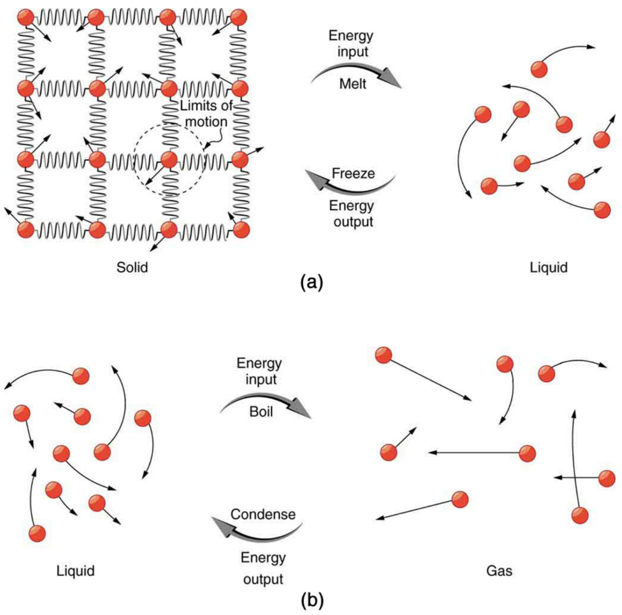
Equation:

$$Q=\mathrm{mL_f}\ (\mathrm{melting/freezing}),$$

Equation:

$$Q=\mathrm{mL_v}\ (\mathrm{vaporization/condensation}),$$

where the latent heat of fusion, L_{f}, and latent heat of vaporization, L_{v}, are material constants that are determined experimentally. See ([link]).



(a) Energy is required to partially overcome the attractive forces between molecules in a solid to form a liquid. That same energy must be removed for freezing to take place. (b) Molecules are separated by large distances when going from liquid to vapor, requiring significant energy to overcome molecular attraction. The same energy must be removed for condensation to take place. There is no temperature change until a phase change is complete.

Latent heat is measured in units of J/kg. Both L_f and L_v depend on the substance, particularly on the strength of its molecular forces as noted earlier. L_f and L_v are collectively called **latent heat coefficients**. They are *latent*, or hidden, because in phase changes, energy enters or leaves a system without causing a temperature change in the system; so, in effect, the energy is hidden. [link] lists representative values of L_f and L_v, together with melting and boiling points.

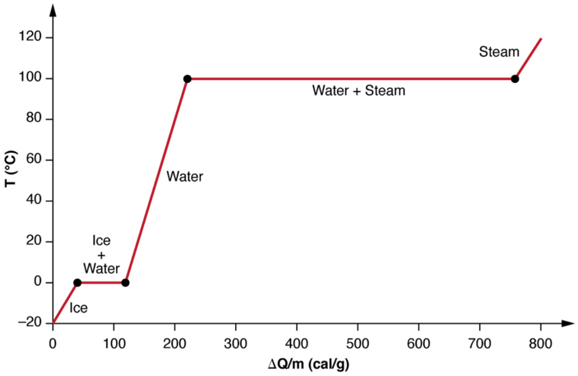
The table shows that significant amounts of energy are involved in phase changes. Let us look, for example, at how much energy is needed to melt a kilogram of ice at $0^\text{º}\text{C}$ to produce a kilogram of water at $0\text{˚}\text{C}$. Using the equation for a change in temperature and the value for water from [link], we find that $Q = mL_\text{f} = (1.0\ \text{kg})(334\ \text{kJ/kg}) = 334\ \text{kJ}$ is the energy to melt a kilogram of ice. This is a lot of energy as it represents the same amount of energy needed to raise the temperature of 1 kg of liquid water from $0^\text{º}\text{C}$ to $79.8^\text{º}\text{C}$. Even more energy is required to vaporize water; it would take 2256 kJ to change 1 kg of liquid water at the normal boiling point ($100^\text{º}\text{C}$ at atmospheric pressure) to steam (water vapor). This example shows that the energy for a phase change is enormous compared to energy associated with temperature changes without a phase change.

		L_f			L_v	
Substance	**Melting point (ºC)**	**kJ/kg**	**kcal/kg**	**Boiling point (°C)**	**kJ/kg**	**kcal/kg**
Helium	−269.7	5.23	1.25	−268.9	20.9	4.99
Hydrogen	−259.3	58.6	14.0	−252.9	452	108
Nitrogen	−210.0	25.5	6.09	−195.8	201	48.0
Oxygen	−218.8	13.8	3.30	−183.0	213	50.9
Ethanol	−114	104	24.9	78.3	854	204
Ammonia	−75		108	−33.4	1370	327
Mercury	−38.9	11.8	2.82	357	272	65.0
Water	0.00	334	79.8	100.0	2256[footnote] At 37.0ºC (body temperature), the heat of vaporization L_v for water is 2430 kJ/kg or 580 kcal/kg	539[footnote] At 37.0ºC (body temperature), the heat of vaporization L_v for water is 2430 kJ/kg or 580 kcal/kg
Sulfur	119	38.1	9.10	444.6	326	77.9
Lead	327	24.5	5.85	1750	871	208
Antimony	631	165	39.4	1440	561	134
Aluminum	660	380	90	2450	11400	2720
Silver	961	88.3	21.1	2193	2336	558
Gold	1063	64.5	15.4	2660	1578	377
Copper	1083	134	32.0	2595	5069	1211
Uranium	1133	84	20	3900	1900	454
Tungsten	3410	184	44	5900	4810	1150

Heats of Fusion and Vaporization [footnote]
Values quoted at the normal melting and boiling temperatures at standard atmospheric pressure (1 atm).

Phase changes can have a tremendous stabilizing effect even on temperatures that are not near the melting and boiling points, because evaporation and condensation (conversion of a gas into a liquid state) occur even at temperatures below the boiling point. Take, for example, the fact that air temperatures in humid climates rarely go above $35.0^{\mathrm{o}}\mathrm{C}$, which is because most heat transfer goes into evaporating water into the air. Similarly, temperatures in humid weather rarely fall below the dew point because enormous heat is released when water vapor condenses.

We examine the effects of phase change more precisely by considering adding heat into a sample of ice at $-20^{\mathrm{o}}\mathrm{C}$ ([link]). The temperature of the ice rises linearly, absorbing heat at a constant rate of $0.50\ \mathrm{cal/g}\cdot^{\mathrm{o}}\mathrm{C}$ until it reaches $0^{\mathrm{o}}\mathrm{C}$. Once at this temperature, the ice begins to melt until all the ice has melted, absorbing 79.8 cal/g of heat. The temperature remains constant at $0^{\mathrm{o}}\mathrm{C}$ during this phase change. Once all the ice has melted, the temperature of the liquid water rises, absorbing heat at a new constant rate of $1.00\ \mathrm{cal/g}\cdot^{\mathrm{o}}\mathrm{C}$. At $100^{\mathrm{o}}\mathrm{C}$, the water begins to boil and the temperature again remains constant while the water absorbs 539 cal/g of heat during this phase change. When all the liquid has become steam vapor, the temperature rises again, absorbing heat at a rate of $0.482\ \mathrm{cal/g}\cdot^{\mathrm{o}}\mathrm{C}$.



A graph of temperature versus energy added. The system is constructed so that no vapor evaporates while ice warms to become liquid water, and so that, when vaporization occurs, the vapor remains in of the system. The long stretches of constant temperature values at $0^{\mathrm{o}}\mathrm{C}$ and $100^{\mathrm{o}}\mathrm{C}$ reflect the large latent heat of melting and vaporization, respectively.

Water can evaporate at temperatures below the boiling point. More energy is required than at the boiling point, because the kinetic energy of water molecules at temperatures below $100^{\mathrm{o}}\mathrm{C}$ is less than that at $100^{\mathrm{o}}\mathrm{C}$, hence less energy is available from random thermal motions. Take, for example, the fact that, at body temperature, perspiration from the skin requires a heat input of 2428 kJ/kg, which is about 10 percent higher than the latent heat of vaporization at $100^{\mathrm{o}}\mathrm{C}$. This heat comes from the skin, and thus provides an effective cooling mechanism in hot weather. High humidity inhibits evaporation, so that body temperature might rise, leaving unevaporated sweat on your brow.

Example:
Calculate Final Temperature from Phase Change: Cooling Soda with Ice Cubes

Three ice cubes are used to chill a soda at $20^{\mathrm{o}}\mathrm{C}$ with mass $m_{\text{soda}} = 0.25\ \text{kg}$. The ice is at $0^{\mathrm{o}}\mathrm{C}$ and each ice cube has a mass of 6.0 g. Assume that the soda is kept in a foam container so that heat loss can be ignored. Assume the soda has the same heat capacity as water. Find the final temperature when all ice has melted.

Strategy

The ice cubes are at the melting temperature of $0^{\mathrm{o}}\mathrm{C}$. Heat is transferred from the soda to the ice for melting. Melting of ice occurs in two steps: first the phase change occurs and solid (ice) transforms into liquid water at the melting temperature, then the temperature of this water rises. Melting yields water at $0^{\mathrm{o}}\mathrm{C}$, so more heat is transferred from the soda to this water until the water plus soda system reaches thermal equilibrium,

Equation:

$$Q_{\text{ice}} = -Q_{\text{soda}}.$$

The heat transferred to the ice is $Q_{\text{ice}} = m_{\text{ice}}L_{\text{f}} + m_{\text{ice}}c_{\text{W}}(T_{\text{f}} - 0^{\mathrm{o}}\mathrm{C})$. The heat given off by the soda is $Q_{\text{soda}} = m_{\text{soda}}c_{\text{W}}(T_{\text{f}} - 20^{\mathrm{o}}\mathrm{C})$. Since no heat is lost, $Q_{\text{ice}} = -Q_{\text{soda}}$, so that

Equation:

$$m_{\text{ice}}L_{\text{f}} + m_{\text{ice}}c_{\text{W}}(T_{\text{f}} - 0^{\mathrm{o}}\mathrm{C}) = -m_{\text{soda}}c_{\text{W}}(T_{\text{f}} - 20^{\mathrm{o}}\mathrm{C}).$$

Bring all terms involving T_{f} on the left-hand-side and all other terms on the right-hand-side. Solve for the unknown quantity T_{f}:

Equation:

$$T_{\text{f}} = \frac{m_{\text{soda}}c_{\text{W}}(20^{\mathrm{o}}\mathrm{C}) - m_{\text{ice}}L_{\text{f}}}{(m_{\text{soda}} + m_{\text{ice}})c_{\text{W}}}.$$

Solution

1. Identify the known quantities. The mass of ice is $m_{\text{ice}} = 3\times6.0\ \text{g} = 0.018\ \text{kg}$ and the mass of soda is $m_{\text{soda}} = 0.25\ \text{kg}$.
2. Calculate the terms in the numerator:

   Equation:

   $$m_{\text{soda}}c_{\text{W}}(20^{\mathrm{o}}\mathrm{C}) = (0.25\ \text{kg})(4186\ \text{J/kg}\ \cdot^{\mathrm{o}}\ \text{C})(20^{\mathrm{o}}\mathrm{C}) = 20{,}930\ \text{J}$$

   and

   Equation:

   $$m_{\text{ice}}L_{\text{f}} = (0.018\ \text{kg})(334{,}000\ \text{J/kg}) {=} 6012\ \text{J}.$$
3. Calculate the denominator:

   Equation:

   $$(m_{\text{soda}} + m_{\text{ice}})c_{\text{W}} = (0.25\ \text{kg} + 0.018\ \text{kg})(4186\ \text{K/(kg}\ \cdot^{\mathrm{o}}\ \text{C}){=}1122\ \text{J/}^{\mathrm{o}}\text{C}.$$
4. Calculate the final temperature:

   Equation:

   $$T_{\text{f}} = \frac{20{,}930\ \text{J} - 6012\ \text{J}}{1122\ \text{J/}^{\mathrm{o}}\text{C}} = 13^{\mathrm{o}}\text{C}.$$

Discussion

This example illustrates the enormous energies involved during a phase change. The mass of ice is about 7 percent the mass of water but leads to a noticeable change in the temperature of soda. Although we assumed that the ice was at the freezing temperature, this is incorrect: the typical temperature is $-6^{\circ}\mathrm{C}$. However, this correction gives a final temperature that is essentially identical to the result we found. Can you explain why?

We have seen that vaporization requires heat transfer to a liquid from the surroundings, so that energy is released by the surroundings. Condensation is the reverse process, increasing the temperature of the surroundings. This increase may seem surprising, since we associate condensation with cold objects—the glass in the figure, for example. However, energy must be removed from the condensing molecules to make a vapor condense. The energy is exactly the same as that required to make the phase change in the other direction, from liquid to vapor, and so it can be calculated from $Q = \mathrm{mL_v}$.

Condensation forms on this glass of iced tea because the temperature of the nearby air is reduced to below the dew point. The rate at which water molecules join together exceeds the rate at which they separate, and so water condenses. Energy is released when the water condenses, speeding the melting of the ice in the glass. (credit: Jenny Downing)

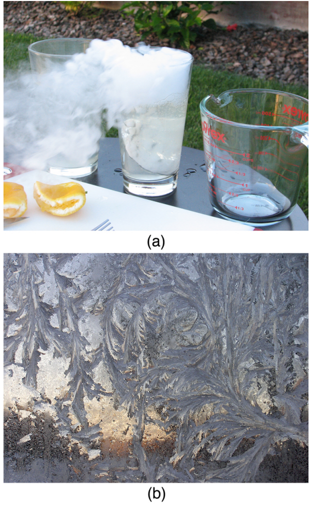
Note:

Real-World Application

Energy is also released when a liquid freezes. This phenomenon is used by fruit growers in Florida to protect oranges when the temperature is close to the freezing point $(0^{\circ}\mathrm{C})$. Growers spray water on the plants in orchards so that the water freezes and heat is released to the growing oranges on the trees. This prevents the temperature inside the orange from dropping below freezing, which would damage the fruit.

The ice on these trees released large amounts of energy when it froze, helping to prevent the temperature of the trees from dropping below 0ºC. Water is intentionally sprayed on orchards to help prevent hard frosts. (credit: Hermann Hammer)

Sublimation is the transition from solid to vapor phase. You may have noticed that snow can disappear into thin air without a trace of liquid water, or the disappearance of ice cubes in a freezer. The reverse is also true: Frost can form on very cold windows without going through the liquid stage. A popular effect is the making of “smoke” from dry ice, which is solid carbon dioxide. Sublimation occurs because the equilibrium vapor pressure of solids is not zero. Certain air fresheners use the sublimation of a solid to inject a perfume into the room. Moth balls are a slightly toxic example of a phenol (an organic compound) that sublimates, while some solids, such as osmium tetroxide, are so toxic that they must be kept in sealed containers to prevent human exposure to their sublimation-produced vapors.



(a)

(b)

Direct transitions between solid and

vapor are common, sometimes useful, and even beautiful. (a) Dry ice sublimates directly to carbon dioxide gas. The visible vapor is made of water droplets. (credit: Windell Oskay) (b) Frost forms patterns on a very cold window, an example of a solid formed directly from a vapor. (credit: Liz West)

All phase transitions involve heat. In the case of direct solid-vapor transitions, the energy required is given by the equation $Q=\mathrm{mL_s}$, where L_s is the **heat of sublimation**, which is the energy required to change 1.00 kg of a substance from the solid phase to the vapor phase. L_s is analogous to L_f and L_v, and its value depends on the substance. Sublimation requires energy input, so that dry ice is an effective coolant, whereas the reverse process (i.e., frosting) releases energy. The amount of energy required for sublimation is of the same order of magnitude as that for other phase transitions.

The material presented in this section and the preceding section allows us to calculate any number of effects related to temperature and phase change. In each case, it is necessary to identify which temperature and phase changes are taking place and then to apply the appropriate equation. Keep in mind that heat transfer and work can cause both temperature and phase changes.

Problem-Solving Strategies for the Effects of Heat Transfer

1. *Examine the situation to determine that there is a change in the temperature or phase. Is there heat transfer into or out of the system?* When the presence or absence of a phase change is not obvious, you may wish to first solve the problem as if there were no phase changes, and examine the temperature change obtained. If it is sufficient to take you past a boiling or melting point, you should then go back and do the problem in steps—temperature change, phase change, subsequent temperature change, and so on.
2. *Identify and list all objects that change temperature and phase.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful.
4. *Make a list of what is given or what can be inferred from the problem as stated (identify the knowns).*
5. *Solve the appropriate equation for the quantity to be determined (the unknown).* If there is a temperature change, the transferred heat depends on the specific heat (see [link]) whereas, for a phase change, the transferred heat depends on the latent heat. See [link].
6. *Substitute the knowns along with their units into the appropriate equation and obtain numerical solutions complete with units.* You will need to do this in steps if there is more than one stage to the process (such as a temperature change followed by a phase change).

7. *Check the answer to see if it is reasonable: Does it make sense?* As an example, be certain that the temperature change does not also cause a phase change that you have not taken into account.

Exercise:
Check Your Understanding

Problem:

Why does snow remain on mountain slopes even when daytime temperatures are higher than the freezing temperature?

Solution:

Snow is formed from ice crystals and thus is the solid phase of water. Because enormous heat is necessary for phase changes, it takes a certain amount of time for this heat to be accumulated from the air, even if the air is above 0°C. The warmer the air is, the faster this heat exchange occurs and the faster the snow melts.

Summary

- Most substances can exist either in solid, liquid, and gas forms, which are referred to as "phases."
- Phase changes occur at fixed temperatures for a given substance at a given pressure, and these temperatures are called boiling and freezing (or melting) points.
- During phase changes, heat absorbed or released is given by:
  Equation:

  $$Q=\text{mL},$$

  where L is the latent heat coefficient.

Conceptual Questions

Exercise:

Problem:

Heat transfer can cause temperature and phase changes. What else can cause these changes?

Exercise:

Problem:

How does the latent heat of fusion of water help slow the decrease of air temperatures, perhaps preventing temperatures from falling significantly below 0°C, in the vicinity of large bodies of water?

Exercise:

Problem: What is the temperature of ice right after it is formed by freezing water?

Exercise:

Problem:

If you place 0°C ice into 0°C water in an insulated container, what will happen? Will some ice melt, will more water freeze, or will neither take place?

Exercise:

Problem:

What effect does condensation on a glass of ice water have on the rate at which the ice melts? Will the condensation speed up the melting process or slow it down?

Exercise:

Problem:

In very humid climates where there are numerous bodies of water, such as in Florida, it is unusual for temperatures to rise above about $35^{\circ}\text{C}(95^{\circ}\text{F})$. In deserts, however, temperatures can rise far above this. Explain how the evaporation of water helps limit high temperatures in humid climates.

Exercise:

Problem:

In winters, it is often warmer in San Francisco than in nearby Sacramento, 150 km inland. In summers, it is nearly always hotter in Sacramento. Explain how the bodies of water surrounding San Francisco moderate its extreme temperatures.

Exercise:

Problem:

Putting a lid on a boiling pot greatly reduces the heat transfer necessary to keep it boiling. Explain why.

Exercise:

Problem:

Freeze-dried foods have been dehydrated in a vacuum. During the process, the food freezes and must be heated to facilitate dehydration. Explain both how the vacuum speeds up dehydration and why the food freezes as a result.

Exercise:

Problem:

When still air cools by radiating at night, it is unusual for temperatures to fall below the dew point. Explain why.

Exercise:

Problem:

In a physics classroom demonstration, an instructor inflates a balloon by mouth and then cools it in liquid nitrogen. When cold, the shrunken balloon has a small amount of light blue liquid in it, as well as some snow-like crystals. As it warms up, the liquid boils, and part of the crystals sublimate, with some crystals lingering for awhile and then producing a liquid. Identify the blue liquid and the two solids in the cold balloon. Justify your identifications using data from [link].

Problems & Exercises

Exercise:

Problem:

How much heat transfer (in kilocalories) is required to thaw a 0.450-kg package of frozen vegetables originally at $0^\text{º}\text{C}$ if their heat of fusion is the same as that of water?

Solution:

35.9 kcal

Exercise:

Problem:

A bag containing $0^\text{º}\text{C}$ ice is much more effective in absorbing energy than one containing the same amount of $0^\text{º}\text{C}$ water.

a. How much heat transfer is necessary to raise the temperature of 0.800 kg of water from $0^\text{º}\text{C}$ to $30.0^\text{º}\text{C}$?
b. How much heat transfer is required to first melt 0.800 kg of $0^\text{º}\text{C}$ ice and then raise its temperature?
c. Explain how your answer supports the contention that the ice is more effective.

Exercise:

Problem:

(a) How much heat transfer is required to raise the temperature of a 0.750-kg aluminum pot containing 2.50 kg of water from $30.0^\text{º}\text{C}$ to the boiling point and then boil away 0.750 kg of water? (b) How long does this take if the rate of heat transfer is 500 W $1\text{ watt} = 1\text{ joule/second}\;\;(1\text{ W} = 1\text{ J/s})$?

Solution:

(a) 591 kcal

(b) $4.94 \times 10^{3}\text{ s}$

Exercise:

Problem:

The formation of condensation on a glass of ice water causes the ice to melt faster than it would otherwise. If 8.00 g of condensation forms on a glass containing both water and 200 g of ice, how many grams of the ice will melt as a result? Assume no other heat transfer occurs.

Exercise:

Problem:

On a trip, you notice that a 3.50-kg bag of ice lasts an average of one day in your cooler. What is the average power in watts entering the ice if it starts at $0^\text{º}\text{C}$ and completely melts to $0^\text{º}\text{C}$ water in exactly one day $1\text{ watt} = 1\text{ joule/second}\;\;\;(1\text{ W} = 1\text{ J/s})$?

Solution:

13.5 W

Exercise:

Problem:

On a certain dry sunny day, a swimming pool’s temperature would rise by $1.50^{\mathrm{o}}\mathrm{C}$ if not for evaporation. What fraction of the water must evaporate to carry away precisely enough energy to keep the temperature constant?

Exercise:

Problem:

(a) How much heat transfer is necessary to raise the temperature of a 0.200-kg piece of ice from $-20.0^{\mathrm{o}}\mathrm{C}$ to $130^{\mathrm{o}}\mathrm{C}$, including the energy needed for phase changes?
(b) How much time is required for each stage, assuming a constant 20.0 kJ/s rate of heat transfer?
(c) Make a graph of temperature versus time for this process.

Solution:

(a) 148 kcal

(b) 0.418 s, 3.34 s, 4.19 s, 22.6 s, 0.456 s

Exercise:

Problem:

In 1986, a gargantuan iceberg broke away from the Ross Ice Shelf in Antarctica. It was approximately a rectangle 160 km long, 40.0 km wide, and 250 m thick.

(a) What is the mass of this iceberg, given that the density of ice is $917\ \mathrm{kg/m^3}$?

(b) How much heat transfer (in joules) is needed to melt it?

(c) How many years would it take sunlight alone to melt ice this thick, if the ice absorbs an average of $100\ \mathrm{W/m^2}$, 12.00 h per day?

Exercise:

Problem:

How many grams of coffee must evaporate from 350 g of coffee in a 100-g glass cup to cool the coffee from $95.0^{\mathrm{o}}\mathrm{C}$ to $45.0^{\mathrm{o}}\mathrm{C}$? You may assume the coffee has the same thermal properties as water and that the average heat of vaporization is 2340 kJ/kg (560 cal/g). (You may neglect the change in mass of the coffee as it cools, which will give you an answer that is slightly larger than correct.)

Solution:

33.0 g

Exercise:

Problem:

(a) It is difficult to extinguish a fire on a crude oil tanker, because each liter of crude oil releases 2.80×10^7 J of energy when burned. To illustrate this difficulty, calculate the number of liters of water that must be expended to absorb the energy released by burning 1.00 L of crude oil, if the water has its temperature raised from 20.0°C to 100°C, it boils, and the resulting steam is raised to 300°C. (b) Discuss additional complications caused by the fact that crude oil has a smaller density than water.

Solution:

(a) 9.67 L

(b) Crude oil is less dense than water, so it floats on top of the water, thereby exposing it to the oxygen in the air, which it uses to burn. Also, if the water is under the oil, it is less efficient in absorbing the heat generated by the oil.

Exercise:

Problem:

The energy released from condensation in thunderstorms can be very large. Calculate the energy released into the atmosphere for a small storm of radius 1 km, assuming that 1.0 cm of rain is precipitated uniformly over this area.

Exercise:

Problem: To help prevent frost damage, 4.00 kg of 0°C water is sprayed onto a fruit tree.

(a) How much heat transfer occurs as the water freezes?

(b) How much would the temperature of the 200-kg tree decrease if this amount of heat transferred from the tree? Take the specific heat to be $3.35 \text{ kJ/kg} \cdot^{\circ} \text{C}$, and assume that no phase change occurs.

Solution:

a) 319 kcal

b) 2.00°C

Exercise:

Problem:

A 0.250-kg aluminum bowl holding 0.800 kg of soup at 25.0°C is placed in a freezer. What is the final temperature if 377 kJ of energy is transferred from the bowl and soup, assuming the soup’s thermal properties are the same as that of water? Explicitly show how you follow the steps in Problem-Solving Strategies for the Effects of Heat Transfer.

Exercise:

Problem:

A 0.0500-kg ice cube at -30.0°C is placed in 0.400 kg of 35.0°C water in a very well-insulated container. What is the final temperature?

Solution:

20.6°C

Exercise:

Problem:

If you pour 0.0100 kg of 20.0°C water onto a 1.20-kg block of ice (which is initially at -15.0°C), what is the final temperature? You may assume that the water cools so rapidly that effects of the surroundings are negligible.

Exercise:

Problem:

Indigenous people sometimes cook in watertight baskets by placing hot rocks into water to bring it to a boil. What mass of 500°C rock must be placed in 4.00 kg of 15.0°C water to bring its temperature to 100°C, if 0.0250 kg of water escapes as vapor from the initial sizzle? You may neglect the effects of the surroundings and take the average specific heat of the rocks to be that of granite.

Solution:

4.38 kg

Exercise:

Problem:

What would be the final temperature of the pan and water in Calculating the Final Temperature When Heat Is Transferred Between Two Bodies: Pouring Cold Water in a Hot Pan if 0.260 kg of water was placed in the pan and 0.0100 kg of the water evaporated immediately, leaving the remainder to come to a common temperature with the pan?

Exercise:

Problem:

In some countries, liquid nitrogen is used on dairy trucks instead of mechanical refrigerators. A 3.00-hour delivery trip requires 200 L of liquid nitrogen, which has a density of $808\ \text{kg/m}^3$.

(a) Calculate the heat transfer necessary to evaporate this amount of liquid nitrogen and raise its temperature to 3.00°C. (Use c_{p} and assume it is constant over the temperature range.) This value is the amount of cooling the liquid nitrogen supplies.

(b) What is this heat transfer rate in kilowatt-hours?

(c) Compare the amount of cooling obtained from melting an identical mass of 0°C ice with that from evaporating the liquid nitrogen.

Solution:

(a) $1.57 \times 10^4\ \text{kcal}$

(b) $18.3\ \text{kW} \cdot \text{h}$

(c) $1.29 \times 10^4\ \text{kcal}$

Exercise:

Problem:

Some gun fanciers make their own bullets, which involves melting and casting the lead slugs. How much heat transfer is needed to raise the temperature and melt 0.500 kg of lead, starting from 25.0°C ?

Glossary

heat of sublimation
the energy required to change a substance from the solid phase to the vapor phase

latent heat coefficient
a physical constant equal to the amount of heat transferred for every 1 kg of a substance during the change in phase of the substance

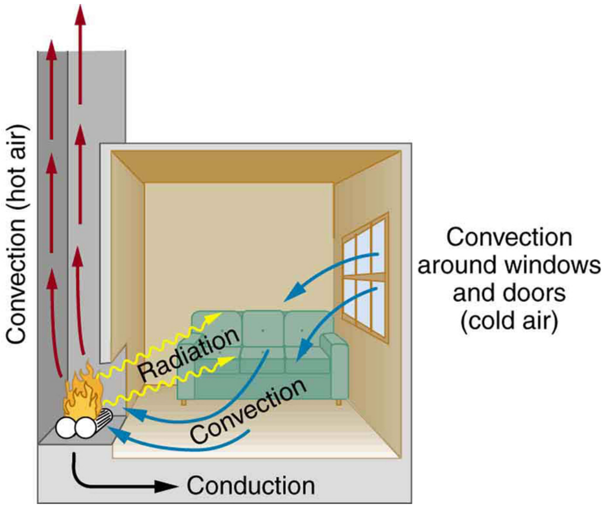
sublimation
the transition from the solid phase to the vapor phase

Heat Transfer Methods

- Discuss the different methods of heat transfer.

Equally as interesting as the effects of heat transfer on a system are the methods by which this occurs. Whenever there is a temperature difference, heat transfer occurs. Heat transfer may occur rapidly, such as through a cooking pan, or slowly, such as through the walls of a picnic ice chest. We can control rates of heat transfer by choosing materials (such as thick wool clothing for the winter), controlling air movement (such as the use of weather stripping around doors), or by choice of color (such as a white roof to reflect summer sunlight). So many processes involve heat transfer, so that it is hard to imagine a situation where no heat transfer occurs. Yet every process involving heat transfer takes place by only three methods:

1. **Conduction** is heat transfer through stationary matter by physical contact. (The matter is stationary on a macroscopic scale—we know there is thermal motion of the atoms and molecules at any temperature above absolute zero.) Heat transferred between the electric burner of a stove and the bottom of a pan is transferred by conduction.
2. **Convection** is the heat transfer by the macroscopic movement of a fluid. This type of transfer takes place in a forced-air furnace and in weather systems, for example.
3. Heat transfer by **radiation** occurs when microwaves, infrared radiation, visible light, or another form of electromagnetic radiation is emitted or absorbed. An obvious example is the warming of the Earth by the Sun. A less obvious example is thermal radiation from the human body.



In a fireplace, heat transfer occurs by all three methods: conduction, convection, and radiation. Radiation is responsible for most of the heat transferred into the room. Heat transfer also occurs through conduction into the room, but at a much slower rate. Heat transfer by convection also occurs through cold air entering the room around windows and hot air leaving the room by rising up the chimney.

We examine these methods in some detail in the three following modules. Each method has unique and interesting characteristics, but all three do have one thing in common: they transfer heat solely because of a temperature difference [link].

Exercise:

Check Your Understanding

Problem:

Name an example from daily life (different from the text) for each mechanism of heat transfer.

Solution:

Conduction: Heat transfers into your hands as you hold a hot cup of coffee.

Convection: Heat transfers as the barista “steams” cold milk to make hot *cocoa*.

Radiation: Reheating a cold cup of coffee in a microwave oven.

Summary

- Heat is transferred by three different methods: conduction, convection, and radiation.

Conceptual Questions

Exercise:

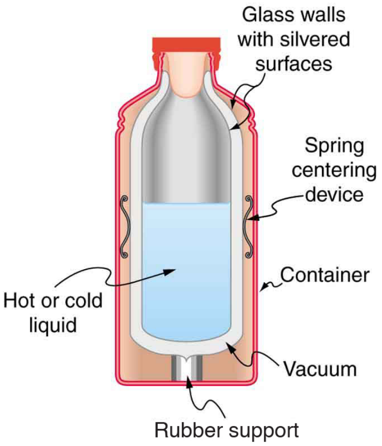
Problem:

What are the main methods of heat transfer from the hot core of Earth to its surface? From Earth’s surface to outer space?

When our bodies get too warm, they respond by sweating and increasing blood circulation to the surface to transfer thermal energy away from the core. What effect will this have on a person in a 40.0°C hot tub?

[link] shows a cut-away drawing of a thermos bottle (also known as a Dewar flask), which is a device designed specifically to slow down all forms of heat transfer. Explain the functions of the various parts, such as the

vacuum, the silvering of the walls, the thin-walled long glass neck, the rubber support, the air layer, and the stopper.



The construction of a thermos bottle is designed to inhibit all methods of heat transfer.

Glossary

conduction
: heat transfer through stationary matter by physical contact

convection
: heat transfer by the macroscopic movement of fluid

radiation
: heat transfer which occurs when microwaves, infrared radiation, visible light, or other electromagnetic radiation is emitted or absorbed

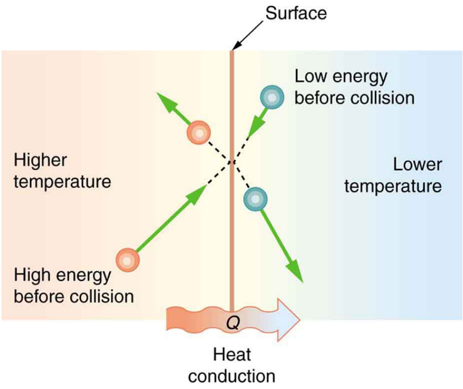
Conduction

- Calculate thermal conductivity.
- Observe conduction of heat in collisions.
- Study thermal conductivities of common substances.

Insulation is used to limit the conduction of heat from the inside to the outside (in winters) and from the outside to the inside (in summers). (credit: Giles Douglas)

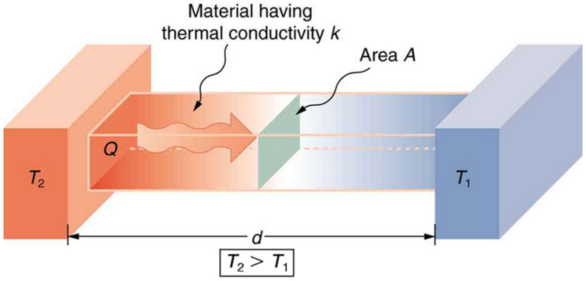
Your feet feel cold as you walk barefoot across the living room carpet in your cold house and then step onto the kitchen tile floor. This result is intriguing, since the carpet and tile floor are both at the same temperature. The different sensation you feel is explained by the different rates of heat transfer: the heat loss during the same time interval is greater for skin in contact with the tiles than with the carpet, so the temperature drop is greater on the tiles.

Some materials conduct thermal energy faster than others. In general, good conductors of electricity (metals like copper, aluminum, gold, and silver) are also good heat conductors, whereas insulators of electricity (wood, plastic, and rubber) are poor heat conductors. [link] shows molecules in two bodies at different temperatures. The (average) kinetic energy of a molecule in the hot body is higher than in the colder body. If two molecules collide, an energy transfer from the molecule with greater kinetic energy to the molecule with less kinetic energy occurs. The cumulative effect from all collisions results in a net flux of heat from the hot body to the colder body. The heat flux thus depends on the temperature difference $\Delta T = T_{\text{hot}} - T_{\text{cold}}$. Therefore, you will get a more severe burn from boiling water than from hot tap water. Conversely, if the temperatures are the same, the net heat transfer rate falls to zero, and equilibrium is achieved. Owing to the fact that the number of collisions increases with increasing area, heat conduction depends on the cross-sectional area. If you touch a cold wall with your palm, your hand cools faster than if you just touch it with your fingertip.



The molecules in two bodies at different temperatures have different average kinetic energies. Collisions occurring at the contact surface tend to transfer energy from high-temperature regions to low-temperature regions. In this illustration, a molecule in the lower temperature region (right side) has low energy before collision, but its energy increases after colliding with the contact surface. In contrast, a molecule in the higher temperature region (left side) has high energy before collision, but its energy decreases after colliding with the contact surface.

A third factor in the mechanism of conduction is the thickness of the material through which heat transfers. The figure below shows a slab of material with different temperatures on either side. Suppose that T_2 is greater than T_1, so that heat is transferred from left to right. Heat transfer from the left side to the right side is accomplished by a series of molecular collisions. The thicker the material, the more time it takes to transfer the same amount of heat. This model explains why thick clothing is warmer than thin clothing in winters, and why Arctic mammals protect themselves with thick blubber.



Heat conduction occurs through any material, represented here by a rectangular bar, whether window glass or walrus blubber. The temperature of the material is T_2 on the left and T_1 on the right, where T_2 is greater than T_1.

The rate of heat transfer by conduction is directly proportional to the surface area A, the temperature difference $T_2 - T_1$, and the substance's conductivity k. The rate of heat transfer is inversely proportional to the thickness d.

Lastly, the heat transfer rate depends on the material properties described by the coefficient of thermal conductivity. All four factors are included in a simple equation that was deduced from and is confirmed by experiments. The **rate of conductive heat transfer** through a slab of material, such as the one in [link], is given by

Equation:

$$\frac{Q}{t} = \frac{\mathrm{kA}(T_2 - T_1)}{d},$$

where Q/t is the rate of heat transfer in watts or kilocalories per second, k is the **thermal conductivity** of the material, A and d are its surface area and thickness, as shown in [link], and $(T_2 - T_1)$ is the temperature difference across the slab. [link] gives representative values of thermal conductivity.

Example:

Calculating Heat Transfer Through Conduction: Conduction Rate Through an Ice Box

A Styrofoam ice box has a total area of $0.950\ \mathrm{m}^2$ and walls with an average thickness of 2.50 cm. The box contains ice, water, and canned beverages at $0^{\circ}\mathrm{C}$. The inside of the box is kept cold by melting ice. How much ice melts in one day if the ice box is kept in the trunk of a car at $35.0^{\circ}\mathrm{C}$?

Strategy

This question involves both heat for a phase change (melting of ice) and the transfer of heat by conduction. To find the amount of ice melted, we must find the net heat transferred. This value can be obtained by calculating the rate of heat transfer by conduction and multiplying by time.

Solution

1. Identify the knowns.

   Equation:

   $$A = 0.950\ \mathrm{m}^2;\ d = 2.50\ \mathrm{cm} = 0.0250\ \mathrm{m};\ T_1 = 0^{\circ}\mathrm{C};\ T_2 = 35.0^{\circ}\mathrm{C},\ t = 1\ \mathrm{day} = 24\ \mathrm{hours} = 86{,}400\ \mathrm{s}.$$

2. Identify the unknowns. We need to solve for the mass of the ice, m. We will also need to solve for the net heat transferred to melt the ice, Q.
3. Determine which equations to use. The rate of heat transfer by conduction is given by

   Equation:

   $$\frac{Q}{t} = \frac{\mathrm{kA}(T_2 - T_1)}{d}.$$

4. The heat is used to melt the ice: $Q = \mathrm{mL_f}$.
5. Insert the known values:

   Equation:

   $$\frac{Q}{t} = \frac{(0.010\ \mathrm{J/s \cdot m \cdot {}^{\circ} C})\left(0.950\ \ \mathrm{m}^2\right)(35.0^{\circ}\mathrm{C} - 0^{\circ}\mathrm{C})}{0.0250\ \mathrm{m}} = 13.3\ \mathrm{J/s}.$$

6. Multiply the rate of heat transfer by the time (1 day $= 86{,}400$ s):
   Equation:

$$Q = (Q/t)t = (13.3\ \text{J/s})(86{,}400\ \text{s}) = 1.15 \times 10^6\ \text{J}.$$

7. Set this equal to the heat transferred to melt the ice: $Q = \mathrm{mL_f}$. Solve for the mass m:
   Equation:

$$m = \frac{Q}{L_\mathrm{f}} = \frac{1.15 \times 10^6\ \text{J}}{334\ \ \times 10^3\ \text{J/kg}} = 3.44\text{kg}.$$

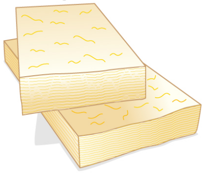
Discussion
The result of 3.44 kg, or about 7.6 lbs, seems about right, based on experience. You might expect to use about a 4 kg (7–10 lb) bag of ice per day. A little extra ice is required if you add any warm food or beverages. Inspecting the conductivities in [link] shows that Styrofoam is a very poor conductor and thus a good insulator. Other good insulators include fiberglass, wool, and goose-down feathers. Like Styrofoam, these all incorporate many small pockets of air, taking advantage of air’s poor thermal conductivity.

Substance	**Thermal conductivity k $(\mathrm{J/s{\cdot}m{\cdot}{}^{o}C})$**
Silver	420
Copper	390
Gold	318
Aluminum	220
Steel iron	80
Steel (stainless)	14
Ice	2.2
Glass (average)	0.84
Concrete brick	0.84
Water	0.6
Fatty tissue (without blood)	0.2
Asbestos	0.16
Plasterboard	0.16
Wood	0.08–0.16

Substance	**Thermal conductivity k $(\mathbf{J/s{\cdot}m{\cdot}^{\circ}C})$**
Snow (dry)	0.10
Cork	0.042
Glass wool	0.042
Wool	0.04
Down feathers	0.025
Air	0.023
Styrofoam	0.010

Thermal Conductivities of Common Substances[footnote]
At temperatures near 0ºC.

A combination of material and thickness is often manipulated to develop good insulators—the smaller the conductivity k and the larger the thickness d, the better. The ratio of d/k will thus be large for a good insulator. The ratio d/k is called the R **factor**. The rate of conductive heat transfer is inversely proportional to R. The larger the value of R, the better the insulation. R factors are most commonly quoted for household insulation, refrigerators, and the like—unfortunately, it is still in non-metric units of $\text{ft}^2{\cdot}°\text{F}{\cdot}\text{h/Btu}$, although the unit usually goes unstated (1 British thermal unit [Btu] is the amount of energy needed to change the temperature of 1.0 lb of water by 1.0 °F). A couple of representative values are an R factor of 11 for 3.5-in-thick fiberglass batts (pieces) of insulation and an R factor of 19 for 6.5-in-thick fiberglass batts. Walls are usually insulated with 3.5-in batts, while ceilings are usually insulated with 6.5-in batts. In cold climates, thicker batts may be used in ceilings and walls.



The fiberglass batt is used for insulation of walls and ceilings to prevent heat transfer between the inside of the building and the outside environment.

Note that in [link], the best thermal conductors—silver, copper, gold, and aluminum—are also the best electrical conductors, again related to the density of free electrons in them. Cooking utensils are typically made from good conductors.

Example:
Calculating the Temperature Difference Maintained by a Heat Transfer: Conduction Through an Aluminum Pan

Water is boiling in an aluminum pan placed on an electrical element on a stovetop. The sauce pan has a bottom that is 0.800 cm thick and 14.0 cm in diameter. The boiling water is evaporating at the rate of 1.00 g/s. What is the temperature difference across (through) the bottom of the pan?

Strategy

Conduction through the aluminum is the primary method of heat transfer here, and so we use the equation for the rate of heat transfer and solve for the temperature difference.

Equation:

$$T_2 - T_1 = \frac{Q}{t}\left(\frac{d}{\mathrm{kA}}\right).$$

Solution

1. Identify the knowns and convert them to the SI units.

   The thickness of the pan, $d = 0.800\text{ cm} = 8.0 \times 10^{-3}\ \text{ m}$, the area of the pan, $A = \pi(0.14/2)^2\ \text{ m}^2 = 1.54 \times 10^{-2}\ \text{ m}^2$, and the thermal conductivity, $k = 220\text{ J/s}\cdot\text{m}\cdot{}^\circ\text{C}$.
2. Calculate the necessary heat of vaporization of 1 g of water:
   Equation:

   $$Q = \mathrm{mL_v} = \left(1.00 \times 10^{-3}\text{ kg}\right)\left(2256 \times 10^3\text{ J/kg}\right) = 2256\text{ J}.$$
3. Calculate the rate of heat transfer given that 1 g of water melts in one second:
   Equation:

   $$Q/t = 2256\text{ J/s or }2.26\text{ kW}.$$
4. Insert the knowns into the equation and solve for the temperature difference:
   Equation:

   $$T_2 - T_1 = \frac{Q}{t}\left(\frac{d}{\mathrm{kA}}\right) = (2256\text{ J/s})\frac{8.00\ \times\ 10^{-3}\text{m}}{(220\text{ J/s}\cdot\text{m}\cdot^{\text{o}}\text{ C})\left(1.54 \times 10^{-2}\text{ m}^2\right)} = 5.33^{\text{o}}\text{C}.$$

Discussion

The value for the heat transfer $Q/t = 2.26\text{kW}$ or 2256 J/s is typical for an electric stove. This value gives a remarkably small temperature difference between the stove and the pan. Consider that the stove burner is red hot while the inside of the pan is nearly 100ºC because of its contact with boiling water. This contact effectively cools the bottom of the pan in spite of its proximity to the very hot stove burner. Aluminum is such a good conductor that it only takes this small temperature difference to produce a heat transfer of 2.26 kW into the pan.

Conduction is caused by the random motion of atoms and molecules. As such, it is an ineffective mechanism for heat transport over macroscopic distances and short time distances. Take, for example, the temperature on the Earth, which would be unbearably cold during the night and extremely hot during the day if heat transport in the atmosphere was to be only through conduction. In another example, car engines would overheat unless there was a more efficient way to remove excess heat from the pistons.

Exercise:
Check Your Understanding

Problem:

How does the rate of heat transfer by conduction change when all spatial dimensions are doubled?

Solution:

Because area is the product of two spatial dimensions, it increases by a factor of four when each dimension is doubled $\left(A_{\text{final}} = (2d)^2 = 4d^2 = 4A_{\text{initial}}\right)$. The distance, however, simply doubles. Because the temperature difference and the coefficient of thermal conductivity are independent of the spatial dimensions, the rate of heat transfer by conduction increases by a factor of four divided by two, or two:

Equation:

$$\left(\frac{Q}{t}\right)_{\text{final}} = \frac{\mathrm{kA}_{\text{final}}(T_2 - T_1)}{d_{\text{final}}} = \frac{k(4\mathrm{A}_{\text{initial}})(T_2 - T_1)}{2\mathrm{d}_{\text{initial}}} = 2\frac{\mathrm{kA}_{\text{initial}}(T_2 - T_1)}{d_{\text{initial}}} = 2\left(\frac{Q}{t}\right)_{\text{initial}}.$$

Summary

- Heat conduction is the transfer of heat between two objects in direct contact with each other.
- The rate of heat transfer Q/t (energy per unit time) is proportional to the temperature difference $T_2 - T_1$ and the contact area A and inversely proportional to the distance d between the objects:
  Equation:

$$\frac{Q}{t} = \frac{\mathrm{kA}(T_2 - T_1)}{d}.$$

Conceptual Questions

Exercise:

Problem:

Some electric stoves have a flat ceramic surface with heating elements hidden beneath. A pot placed over a heating element will be heated, while it is safe to touch the surface only a few centimeters away. Why is ceramic, with a conductivity less than that of a metal but greater than that of a good insulator, an ideal choice for the stove top?

Exercise:

Problem:

Loose-fitting white clothing covering most of the body is ideal for desert dwellers, both in the hot Sun and during cold evenings. Explain how such clothing is advantageous during both day and night.

A jellabiya is worn by many men in Egypt. (credit: Zerida)

Problems & Exercises

Exercise:

Problem:

(a) Calculate the rate of heat conduction through house walls that are 13.0 cm thick and that have an average thermal conductivity twice that of glass wool. Assume there are no windows or doors. The surface area of the walls is $120\ \text{m}^2$ and their inside surface is at 18.0°C, while their outside surface is at 5.00°C. (b) How many 1-kW room heaters would be needed to balance the heat transfer due to conduction?

Solution:

(a) 1.01×10^3 W

(b) One

Exercise:

Problem:

The rate of heat conduction out of a window on a winter day is rapid enough to chill the air next to it. To see just how rapidly the windows transfer heat by conduction, calculate the rate of conduction in watts through a 3.00-m^2 window that is $0.635\ \text{cm}$ thick (1/4 in) if the temperatures of the inner and outer surfaces are 5.00°C and -10.0°C, respectively. This rapid rate will not be maintained—the inner surface will cool, and even result in frost formation.

Exercise:

Problem:

Calculate the rate of heat conduction out of the human body, assuming that the core internal temperature is 37.0°C, the skin temperature is 34.0°C, the thickness of the tissues between averages $1.00\ \text{cm}$, and the surface area is $1.40\ \text{m}^2$.

Solution:

84.0 W

Exercise:

Problem:

Suppose you stand with one foot on ceramic flooring and one foot on a wool carpet, making contact over an area of 80.0 cm^2 with each foot. Both the ceramic and the carpet are 2.00 cm thick and are 10.0°C on their bottom sides. At what rate must heat transfer occur from each foot to keep the top of the ceramic and carpet at 33.0°C?

Exercise:

Problem:

A man consumes 3000 kcal of food in one day, converting most of it to maintain body temperature. If he loses half this energy by evaporating water (through breathing and sweating), how many kilograms of water evaporate?

Solution:

2.59 kg

Exercise:

Problem:

(a) A firewalker runs across a bed of hot coals without sustaining burns. Calculate the heat transferred by conduction into the sole of one foot of a firewalker given that the bottom of the foot is a 3.00-mm-thick callus with a conductivity at the low end of the range for wood and its density is 300 kg/m^3. The area of contact is 25.0 cm^2, the temperature of the coals is 700°C, and the time in contact is 1.00 s.

(b) What temperature increase is produced in the 25.0 cm^3 of tissue affected?

(c) What effect do you think this will have on the tissue, keeping in mind that a callus is made of dead cells?

Exercise:

Problem:

(a) What is the rate of heat conduction through the 3.00-cm-thick fur of a large animal having a 1.40-m^2 surface area? Assume that the animal’s skin temperature is 32.0°C, that the air temperature is -5.00°C, and that fur has the same thermal conductivity as air. (b) What food intake will the animal need in one day to replace this heat transfer?

Solution:

(a) 39.7 W

(b) 820 kcal

Exercise:

Problem:

A walrus transfers energy by conduction through its blubber at the rate of 150 W when immersed in -1.00°C water. The walrus's internal core temperature is 37.0°C, and it has a surface area of $2.00\ \text{m}^2$. What is the average thickness of its blubber, which has the conductivity of fatty tissues without blood?

Walrus on ice. (credit: Captain Budd Christman, NOAA Corps)

Exercise:

Problem:

Compare the rate of heat conduction through a 13.0-cm-thick wall that has an area of $10.0\ \text{m}^2$ and a thermal conductivity twice that of glass wool with the rate of heat conduction through a window that is 0.750 cm thick and that has an area of $2.00\ \text{m}^2$, assuming the same temperature difference across each.

Solution:

35 to 1, window to wall

Exercise:

Problem:

Suppose a person is covered head to foot by wool clothing with average thickness of 2.00 cm and is transferring energy by conduction through the clothing at the rate of 50.0 W. What is the temperature difference across the clothing, given the surface area is $1.40\ \text{m}^2$?

Exercise:

Problem:

Some stove tops are smooth ceramic for easy cleaning. If the ceramic is 0.600 cm thick and heat conduction occurs through the same area and at the same rate as computed in [link], what is the temperature difference across it? Ceramic has the same thermal conductivity as glass and brick.

Solution:

$1.05 \times 10^3\ \text{K}$

Exercise:

Problem:

One easy way to reduce heating (and cooling) costs is to add extra insulation in the attic of a house. Suppose the house already had 15 cm of fiberglass insulation in the attic and in all the exterior surfaces. If you added an extra 8.0 cm of fiberglass to the attic, then by what percentage would the heating cost of the house drop? Take the single story house to be of dimensions 10 m by 15 m by 3.0 m. Ignore air infiltration and heat loss through windows and doors.

Exercise:

Problem:

(a) Calculate the rate of heat conduction through a double-paned window that has a 1.50-m^2 area and is made of two panes of 0.800-cm-thick glass separated by a 1.00-cm air gap. The inside surface temperature is 15.0°C, while that on the outside is -10.0°C. (Hint: There are identical temperature drops across the two glass panes. First find these and then the temperature drop across the air gap. This problem ignores the increased heat transfer in the air gap due to convection.)

(b) Calculate the rate of heat conduction through a 1.60-cm-thick window of the same area and with the same temperatures. Compare your answer with that for part (a).

Solution:

(a) 83 W

(b) 24 times that of a double pane window.

Exercise:

Problem:

Many decisions are made on the basis of the payback period: the time it will take through savings to equal the capital cost of an investment. Acceptable payback times depend upon the business or philosophy one has. (For some industries, a payback period is as small as two years.) Suppose you wish to install the extra insulation in [link]. If energy cost $1.00 per million joules and the insulation was $4.00 per square meter, then calculate the simple payback time. Take the average ΔT for the 120 day heating season to be 15.0°C.

Exercise:

Problem:

For the human body, what is the rate of heat transfer by conduction through the body's tissue with the following conditions: the tissue thickness is 3.00 cm, the change in temperature is 2.00°C, and the skin area is $1.50\ \text{m}^2$. How does this compare with the average heat transfer rate to the body resulting from an energy intake of about 2400 kcal per day? (No exercise is included.)

Solution:

20.0 W, 17.2% of 2400 kcal per day

Glossary

R factor
: the ratio of thickness to the conductivity of a material

rate of conductive heat transfer
: rate of heat transfer from one material to another

thermal conductivity
the property of a material’s ability to conduct heat

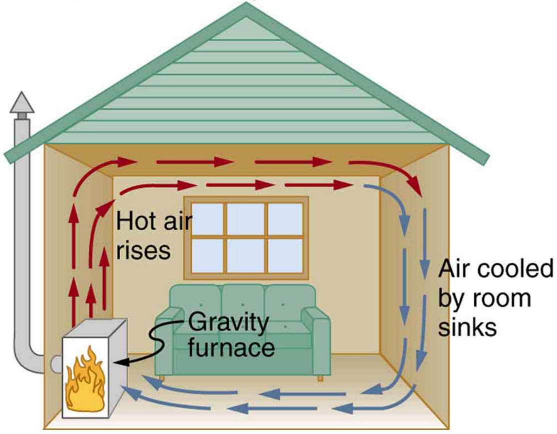
Convection

- Discuss the method of heat transfer by convection.

Convection is driven by large-scale flow of matter. In the case of Earth, the atmospheric circulation is caused by the flow of hot air from the tropics to the poles, and the flow of cold air from the poles toward the tropics. (Note that Earth’s rotation causes the observed easterly flow of air in the northern hemisphere). Car engines are kept cool by the flow of water in the cooling system, with the water pump maintaining a flow of cool water to the pistons. The circulatory system is used the body: when the body overheats, the blood vessels in the skin expand (dilate), which increases the blood flow to the skin where it can be cooled by sweating. These vessels become smaller when it is cold outside and larger when it is hot (so more fluid flows, and more energy is transferred).

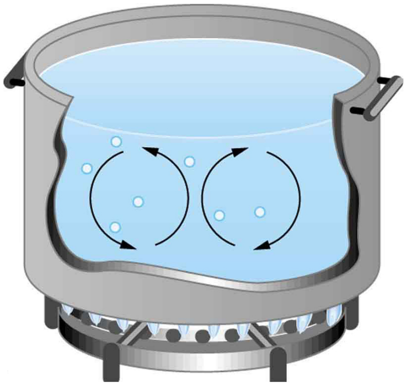
The body also loses a significant fraction of its heat through the breathing process.

While convection is usually more complicated than conduction, we can describe convection and do some straightforward, realistic calculations of its effects. Natural convection is driven by buoyant forces: hot air rises because density decreases as temperature increases. The house in [link] is kept warm in this manner, as is the pot of water on the stove in [link]. Ocean currents and large-scale atmospheric circulation transfer energy from one part of the globe to another. Both are examples of natural convection.



Air heated by the so-called

gravity furnace expands and rises, forming a convective loop that transfers energy to other parts of the room. As the air is cooled at the ceiling and outside walls, it contracts, eventually becoming denser than room air and sinking to the floor. A properly designed heating system using natural convection, like this one, can be quite efficient in uniformly heating a home.



Convection plays an important role in heat transfer inside this pot of water. Once conducted to the inside, heat transfer to other parts of the pot is mostly by convection. The hotter water expands, decreases in density, and rises to transfer heat to other regions of the water, while colder water sinks to the bottom. This process keeps repeating.

Note:

Take-Home Experiment: Convection Rolls in a Heated Pan

Take two small pots of water and use an eye dropper to place a drop of food coloring near the bottom of each. Leave one on a bench top and heat the other over a stovetop. Watch how the color spreads and how long it takes the color to reach the top. Watch how convective loops form.

Example:

Calculating Heat Transfer by Convection: Convection of Air Through the Walls of a House

Most houses are not airtight: air goes in and out around doors and windows, through cracks and crevices, following wiring to switches and outlets, and so on. The air in a typical house is completely replaced in less than an hour. Suppose that a moderately-sized house has inside dimensions $12.0\text{m} \times 18.0\text{m} \times 3.00\text{m}$ high, and that all air is replaced in 30.0 min. Calculate the heat transfer per unit time in watts needed to warm the incoming cold air by 10.0ºC, thus replacing the heat transferred by convection alone.

Strategy

Heat is used to raise the temperature of air so that $Q = \text{mc}\Delta T$. The rate of heat transfer is then Q/t, where t is the time for air turnover. We are given that ΔT is 10.0ºC, but we must still find values for the mass of air and its specific heat before we can calculate Q. The specific heat of air is a weighted average of the specific heats of nitrogen and oxygen, which gives $c = c_{\text{p}} \cong 1000 \text{ J/kg} \cdot^{\text{o}} \text{C}$ from [link] (note that the specific heat at constant pressure must be used for this process).

Solution

1. Determine the mass of air from its density and the given volume of the house. The density is given from the density ρ and the volume
   Equation:
   $$m = \rho\text{V} = \left(1.29 \text{ kg/m}^3\right)(12.0 \text{ m} \times 18.0 \text{ m} \times 3.00 \text{ m}) = 836 \text{ kg}.$$
2. Calculate the heat transferred from the change in air temperature: $Q = \text{mc}\Delta T$ so that
   Equation:
   $$Q = (836 \text{ kg})(1000 \text{ J/kg} \cdot^{\text{o}} \text{C})(10.0^{\text{o}}\text{C}) = 8.36 \times 10^6 \text{ J}.$$
3. Calculate the heat transfer from the heat Q and the turnover time t. Since air is turned over in $t = 0.500 \text{ h} = 1800 \text{ s}$, the heat transferred per unit time is
   Equation:
   $$\frac{Q}{t} = \frac{8.36 \times 10^6 \text{ J}}{1800 \text{ s}} = 4.64 \text{ kW}.$$

Discussion

This rate of heat transfer is equal to the power consumed by about forty-six 100-W light bulbs. Newly constructed homes are designed for a turnover time of 2 hours or more, rather than 30 minutes for the house of this example. Weather stripping, caulking, and improved window seals are commonly employed. More extreme measures are sometimes taken in very cold (or hot) climates to achieve a tight standard of more than 6 hours for one air turnover. Still longer turnover times are unhealthy, because a minimum amount of fresh air is necessary to supply oxygen for breathing and to dilute household pollutants. The term used for the process by which outside air leaks into the house from cracks around windows, doors, and the foundation is called “air infiltration.”

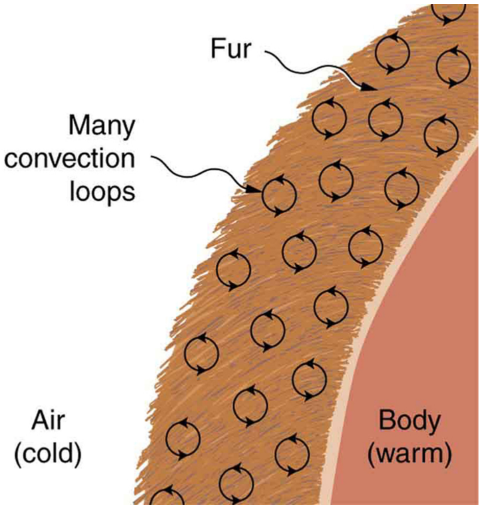
A cold wind is much more chilling than still cold air, because convection combines with conduction in the body to increase the rate at which energy is transferred away from the body. The table below gives approximate wind-chill factors, which are the temperatures of still air that produce the same rate of cooling as air of a given temperature and speed. Wind-chill factors are a dramatic reminder of convection’s ability to transfer heat faster than conduction. For example, a 15.0 m/s wind at 0°C has the chilling equivalent of still air at about -18°C.

Moving air temperature	**Wind speed (m/s)**				
(ºC)	**2**	**5**	**10**	**15**	**20**
5	3	−1	−8	−10	−12
2	0	−7	−12	−16	−18
0	−2	−9	−15	−18	−20

Moving air temperature	**Wind speed (m/s)**				
−5	−7	−15	−22	−26	−29
−10	−12	−21	−29	−34	−36
−20	−23	−34	−44	−50	−52
−40	−44	−59	−73	−82	−84

Wind-Chill Factors

Although air can transfer heat rapidly by convection, it is a poor conductor and thus a good insulator. The amount of available space for airflow determines whether air acts as an insulator or conductor. The space between the inside and outside walls of a house, for example, is about 9 cm (3.5 in)—large enough for convection to work effectively. The addition of wall insulation prevents airflow, so heat loss (or gain) is decreased. Similarly, the gap between the two panes of a double-paned window is about 1 cm, which prevents convection and takes advantage of air’s low conductivity to prevent greater loss. Fur, fiber, and fiberglass also take advantage of the low conductivity of air by trapping it in spaces too small to support convection, as shown in the figure. Fur and feathers are lightweight and thus ideal for the protection of animals.



Fur is filled with air, breaking it up into many small pockets. Convection is very slow here, because the loops are so small. The low conductivity of air makes fur a very good lightweight insulator.

Some interesting phenomena happen *when convection is accompanied by a phase change*. It allows us to cool off by sweating, even if the temperature of the surrounding air exceeds body temperature. Heat from the skin is required for sweat to evaporate from the skin, but without air flow, the air becomes saturated and evaporation stops. Air flow caused by convection replaces the saturated air by dry air and evaporation continues.

Example:
Calculate the Flow of Mass during Convection: Sweat-Heat Transfer away from the Body

The average person produces heat at the rate of about 120 W when at rest. At what rate must water evaporate from the body to get rid of all this energy? (This evaporation might occur when a person is sitting in the shade and surrounding temperatures are the same as skin temperature, eliminating heat transfer by other methods.)

Strategy

Energy is needed for a phase change ($Q = \text{mL}_\text{v}$). Thus, the energy loss per unit time is

Equation:

$$\frac{Q}{t} = \frac{\text{mL}_\text{v}}{t} = 120\ \ \text{W} = 120\ \text{J/s}.$$

We divide both sides of the equation by L_v to find that the mass evaporated per unit time is

Equation:

$$\frac{m}{t} = \frac{120\ \text{J/s}}{L_\text{v}}.$$

Solution

(1) Insert the value of the latent heat from [link], $L_\text{v} = 2430\ \text{kJ/kg} = 2430\ \text{J/g}$. This yields

Equation:

$$\frac{m}{t} = \frac{120\ \text{J/s}}{2430\ \text{J/g}} = 0.0494\ \text{g/s} = 2.96\ \text{g/min}.$$

Discussion

Evaporating about 3 g/min seems reasonable. This would be about 180 g (about 7 oz) per hour. If the air is very dry, the sweat may evaporate without even being noticed. A significant amount of evaporation also takes place in the lungs and breathing passages.

Another important example of the combination of phase change and convection occurs when water evaporates from the oceans. Heat is removed from the ocean when water evaporates. If the water vapor condenses in liquid droplets as clouds form, heat is released in the atmosphere. Thus, there is an overall transfer of heat from the ocean to the atmosphere. This process is the driving power behind thunderheads, those great cumulus clouds that rise as much as 20.0 km into the stratosphere. Water vapor carried in by convection condenses, releasing tremendous amounts of energy. This energy causes the air to expand and rise, where it is colder. More condensation occurs in these colder regions, which in turn drives the cloud even higher. Such a mechanism is called positive feedback, since the process reinforces and accelerates itself. These systems sometimes produce violent storms, with lightning and hail, and constitute the mechanism driving hurricanes.

Cumulus clouds are caused by water vapor that rises because of convection. The rise of clouds is driven by a positive feedback mechanism. (credit: Mike Love)

Convection accompanied by a phase change releases the energy needed to drive this thunderhead into the stratosphere. (credit: Gerardo García Moretti)

The phase change that occurs when this iceberg melts involves tremendous heat transfer. (credit: Dominic Alves)

The movement of icebergs is another example of convection accompanied by a phase change. Suppose an iceberg drifts from Greenland into warmer Atlantic waters. Heat is removed from the warm ocean water when the ice melts and heat is released to the land mass when the iceberg forms on Greenland.

Exercise:

Check Your Understanding

Problem: Explain why using a fan in the summer feels refreshing!

Solution:

Using a fan increases the flow of air: warm air near your body is replaced by cooler air from elsewhere. Convection increases the rate of heat transfer so that moving air “feels” cooler than still air.

Summary

- Convection is heat transfer by the macroscopic movement of mass. Convection can be natural or forced and generally transfers thermal energy faster than conduction. [link] gives wind-chill factors, indicating that moving air has the same chilling effect of much colder stationary air. *Convection that occurs along with a phase change* can transfer energy from cold regions to warm ones.

Conceptual Questions

Exercise:

Problem:

One way to make a fireplace more energy efficient is to have an external air supply for the combustion of its fuel. Another is to have room air circulate around the outside of the fire box and back into the room. Detail the methods of heat transfer involved in each.

Exercise:

Problem:

On cold, clear nights horses will sleep under the cover of large trees. How does this help them keep warm?

Problems & Exercises

Exercise:

Problem:

At what wind speed does -10^{o}C air cause the same chill factor as still air at -29^{o}C?

Solution:

10 m/s

Exercise:

Problem:

At what temperature does still air cause the same chill factor as -5^{o}C air moving at 15 m/s?

Exercise:

Problem:

The “steam” above a freshly made cup of instant coffee is really water vapor droplets condensing after evaporating from the hot coffee. What is the final temperature of 250 g of hot coffee initially at 90.0°C if 2.00 g evaporates from it? The coffee is in a Styrofoam cup, so other methods of heat transfer can be neglected.

Solution:

85.7°C

Exercise:

Problem:

(a) How many kilograms of water must evaporate from a 60.0-kg woman to lower her body temperature by 0.750°C?

(b) Is this a reasonable amount of water to evaporate in the form of perspiration, assuming the relative humidity of the surrounding air is low?

Exercise:

Problem:

On a hot dry day, evaporation from a lake has just enough heat transfer to balance the $1.00\ \text{kW/m}^2$ of incoming heat from the Sun. What mass of water evaporates in 1.00 h from each square meter? Explicitly show how you follow the steps in the Problem-Solving Strategies for the Effects of Heat Transfer.

Solution:

1.48 kg

Exercise:

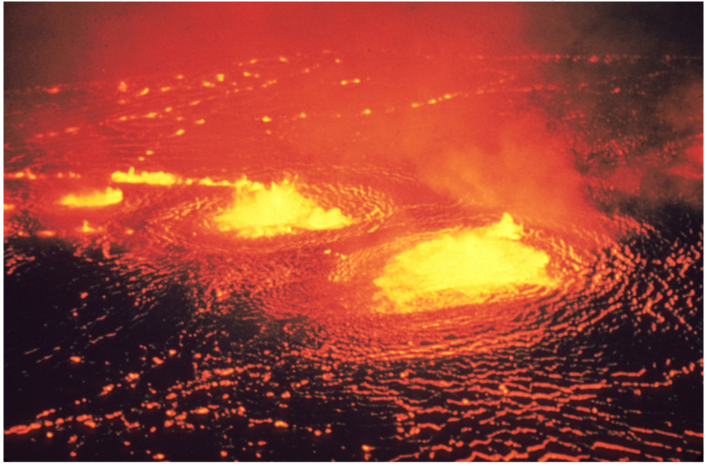
Problem:

One winter day, the climate control system of a large university classroom building malfunctions. As a result, $500\ \mathrm{m}^3$ of excess cold air is brought in each minute. At what rate in kilowatts must heat transfer occur to warm this air by $10.0^{\circ}\mathrm{C}$ (that is, to bring the air to room temperature)?

Exercise:

Problem:

The Kilauea volcano in Hawaii is the world’s most active, disgorging about $5 \times 10^5\ \mathrm{m}^3$ of $1200^{\circ}\mathrm{C}$ lava per day. What is the rate of heat transfer out of Earth by convection if this lava has a density of $2700\ \mathrm{kg/m}^3$ and eventually cools to $30^{\circ}\mathrm{C}$? Assume that the specific heat of lava is the same as that of granite.



Lava flow on Kilauea volcano in Hawaii. (credit: J. P. Eaton, U.S. Geological Survey)

Solution:

$2 \times 10^4\ \mathrm{MW}$

Exercise:

Problem:

During heavy exercise, the body pumps 2.00 L of blood per minute to the surface, where it is cooled by 2.00°C. What is the rate of heat transfer from this forced convection alone, assuming blood has the same specific heat as water and its density is $1050\ \text{kg/m}^3$?

Exercise:

Problem:

A person inhales and exhales 2.00 L of 37.0°C air, evaporating 4.00×10^{-2} g of water from the lungs and breathing passages with each breath.

(a) How much heat transfer occurs due to evaporation in each breath?

(b) What is the rate of heat transfer in watts if the person is breathing at a moderate rate of 18.0 breaths per minute?

(c) If the inhaled air had a temperature of 20.0°C, what is the rate of heat transfer for warming the air?

(d) Discuss the total rate of heat transfer as it relates to typical metabolic rates. Will this breathing be a major form of heat transfer for this person?

Solution:

(a) 97.2 J

(b) 29.2 W

(c) 9.49 W

(d) The total rate of heat loss would be $29.2\ \text{W} + 9.49\ \text{W} = 38.7\ \text{W}$. While sleeping, our body consumes 83 W of power, while sitting it consumes 120 to 210 W. Therefore, the total rate of heat loss from breathing will not be a major form of heat loss for this person.

Exercise:

Problem:

A glass coffee pot has a circular bottom with a 9.00-cm diameter in contact with a heating element that keeps the coffee warm with a continuous heat transfer rate of 50.0 W

(a) What is the temperature of the bottom of the pot, if it is 3.00 mm thick and the inside temperature is 60.0°C?

(b) If the temperature of the coffee remains constant and all of the heat transfer is removed by evaporation, how many grams per minute evaporate? Take the heat of vaporization to be 2340 kJ/kg.

Radiation

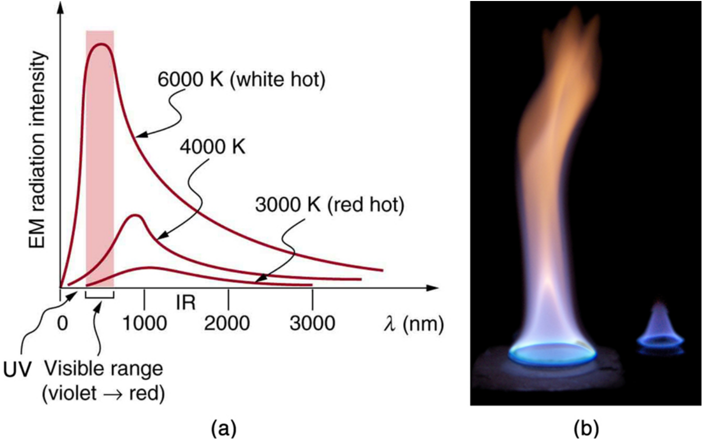
- Discuss heat transfer by radiation.
- Explain the power of different materials.

You can feel the heat transfer from a fire and from the Sun. Similarly, you can sometimes tell that the oven is hot without touching its door or looking inside—it may just warm you as you walk by. The space between the Earth and the Sun is largely empty, without any possibility of heat transfer by convection or conduction. In these examples, heat is transferred by radiation. That is, the hot body emits electromagnetic waves that are absorbed by our skin: no medium is required for electromagnetic waves to propagate. Different names are used for electromagnetic waves of different wavelengths: radio waves, microwaves, infrared **radiation**, visible light, ultraviolet radiation, X-rays, and gamma rays.

Most of the heat transfer from this fire to the observers is through infrared radiation. The visible light, although dramatic, transfers relatively little thermal energy. Convection transfers energy away from the observers as hot air rises, while conduction is negligibly slow here. Skin is very sensitive to infrared radiation, so that you can sense the presence of a fire without looking at it directly. (credit: Daniel X. O’Neil)

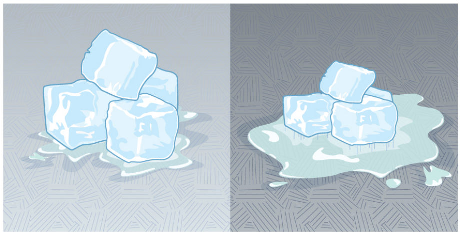
The energy of electromagnetic radiation depends on the wavelength (color) and varies over a wide range: a smaller wavelength (or higher frequency) corresponds to a higher energy. Because more heat is radiated at higher temperatures, a temperature change is accompanied by a color change. Take, for example, an electrical element on a stove, which glows from red to orange, while the higher-temperature steel in a blast furnace glows from yellow to white. The radiation you feel is mostly infrared, which corresponds to a lower temperature than that of the electrical element and the steel. The radiated energy depends on its intensity, which is represented in the figure below by the height of the distribution.

Electromagnetic Waves explains more about the electromagnetic spectrum and Introduction to Quantum Physics discusses how the decrease in wavelength corresponds to an increase in energy.



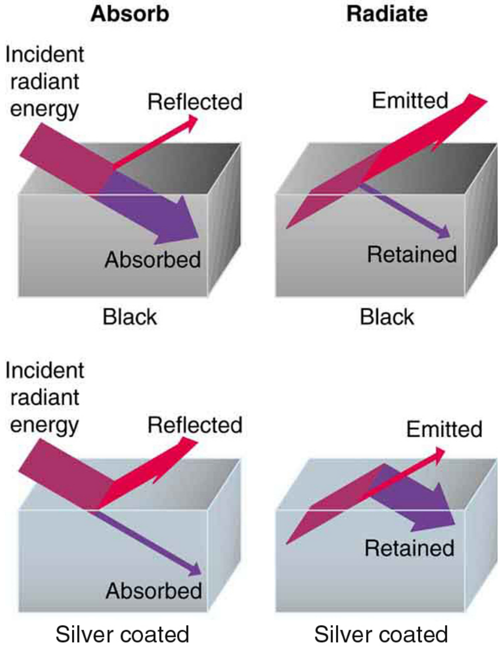
(a) A graph of the spectra of electromagnetic waves emitted from an ideal radiator at three different temperatures. The intensity or rate of radiation emission increases dramatically with temperature, and the spectrum shifts toward the visible and ultraviolet parts of the spectrum. The shaded portion denotes the visible part of the spectrum. It is apparent that the shift toward the ultraviolet with temperature makes the visible appearance shift from red to white to blue as temperature increases. (b) Note the variations in color corresponding to variations in flame temperature. (credit: Tuohirulla)

All objects absorb and emit electromagnetic radiation. The rate of heat transfer by radiation is largely determined by the color of the object. Black is the most effective, and white is the least effective. People living in hot climates generally avoid wearing black clothing, for instance (see [link]). Similarly, black asphalt in a parking lot will be hotter than adjacent gray sidewalk on a summer day, because black absorbs better than gray. The reverse is also true—black radiates better than gray. Thus, on a clear summer night, the asphalt will be colder than the gray sidewalk, because black radiates the energy more rapidly than gray. An *ideal radiator* is the same color as an *ideal absorber*, and captures all the radiation that falls on it. In contrast, white is a poor absorber and is also a poor radiator. A white object reflects all radiation, like a mirror. (A perfect, polished white surface is mirror-like in appearance, and a crushed mirror looks white.)



This illustration shows that the darker pavement is hotter than the lighter pavement (much more of the ice on the right has melted), although both have been in the sunlight for the same time. The thermal conductivities of the pavements are the same.

Gray objects have a uniform ability to absorb all parts of the electromagnetic spectrum. Colored objects behave in similar but more complex ways, which gives them a particular color in the visible range and may make them special in other ranges of the nonvisible spectrum. Take, for example, the strong absorption of infrared radiation by the skin, which allows us to be very sensitive to it.



A black object is a good absorber and a good radiator, while a white (or silver) object is a poor absorber and a poor radiator. It is as if radiation from the inside is reflected back into the silver object, whereas radiation from the inside of the black object is “absorbed” when it hits the surface and finds itself on the outside and is strongly emitted.

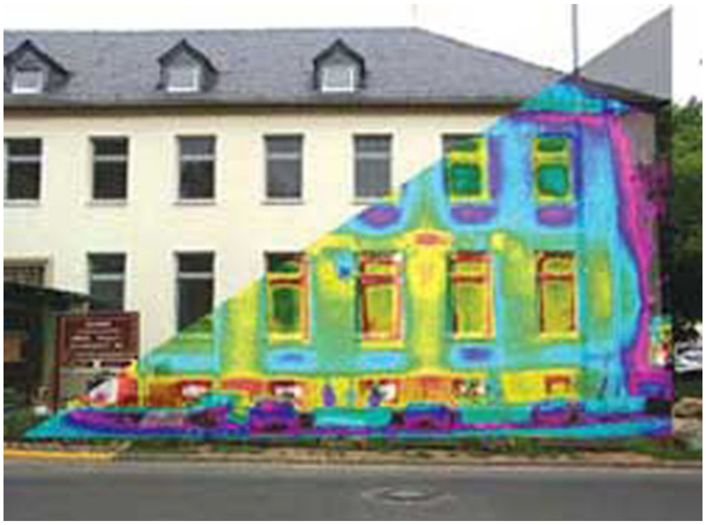
The rate of heat transfer by emitted radiation is determined by the **Stefan-Boltzmann law of radiation**:

Equation:

$$\frac{Q}{t} = \sigma e A T^4,$$

where $\sigma = 5.67 \times 10^{-8}\ \mathrm{J/s \cdot m^2 \cdot K^4}$ is the Stefan-Boltzmann constant, A is the surface area of the object, and T is its absolute temperature in kelvin. The symbol e stands for the **emissivity** of the object, which is a measure of how well it radiates. An ideal jet-black (or black body) radiator has $e = 1$, whereas a perfect reflector has $e = 0$. Real objects fall between these two values. Take, for example, tungsten light bulb filaments which have an e of about 0.5, and carbon black (a material used in printer toner), which has the (greatest known) emissivity of about 0.99.

The radiation rate is directly proportional to the *fourth power* of the absolute temperature—a remarkably strong temperature dependence. Furthermore, the radiated heat is proportional to the surface area of the object. If you knock apart the coals of a fire, there is a noticeable increase in radiation due to an increase in radiating surface area.



A thermograph of part of a building shows temperature variations, indicating where heat transfer to the outside is most severe. Windows are a major region of heat transfer to the outside of homes. (credit: U.S. Army)

Skin is a remarkably good absorber and emitter of infrared radiation, having an emissivity of 0.97 in the infrared spectrum. Thus, we are all nearly (jet) black in the infrared, in spite of the obvious variations in skin color. This high infrared emissivity is why we can so easily feel radiation on our skin. It is also the basis for the use of night scopes used by law enforcement and the military to detect human beings. Even small temperature variations can be detected because of the T^4 dependence. Images, called *thermographs*, can be used medically to detect regions of abnormally high temperature in the body, perhaps indicative of disease. Similar techniques can be used to detect heat leaks in homes [link], optimize performance of blast furnaces, improve comfort levels in work environments, and even remotely map the Earth's temperature profile.

All objects emit and absorb radiation. The *net* rate of heat transfer by radiation (absorption minus emission) is related to both the temperature of the object and the temperature of its surroundings. Assuming that an object with a temperature T_1 is surrounded by an environment with uniform temperature T_2, the **net rate of heat transfer by radiation** is

Equation:

$$\frac{Q_{\text{net}}}{t} = \sigma e A\left(T_2^4 - T_1^4\right),$$

where e is the emissivity of the object alone. In other words, it does not matter whether the surroundings are white, gray, or black; the balance of radiation into and out of the object depends on how well it emits and absorbs radiation. When $T_2 > T_1$, the quantity Q_{net}/t is positive; that is, the net heat transfer is from hot to cold.

Note:

Take-Home Experiment: Temperature in the Sun

Place a thermometer out in the sunshine and shield it from direct sunlight using an aluminum foil. What is the reading? Now remove the shield, and note what the thermometer reads. Take a handkerchief soaked in nail polish remover, wrap it around the thermometer and place it in the sunshine. What does the thermometer read?

Example:

Calculate the Net Heat Transfer of a Person: Heat Transfer by Radiation

What is the rate of heat transfer by radiation, with an unclothed person standing in a dark room whose ambient temperature is 22.0^{o}C. The person has a normal skin temperature of 33.0^{o}C and a surface area of $1.50\ \text{m}^2$. The emissivity of skin is 0.97 in the infrared, where the radiation takes place.

Strategy

We can solve this by using the equation for the rate of radiative heat transfer.

Solution

Insert the temperatures values $T_2 = 295\text{ K}$ and $T_1 = 306\text{ K}$, so that

Equation:

$$\frac{Q}{t}{=}\sigma eA\left(T_2^4 - T_1^4\right)$$

Equation:

$$= \left(5.67 \times 10^{-8}\text{ J/s} \cdot \text{ m}^2 \cdot \text{ K}^4\right)(0.97)\left(1.50\text{ m}^2\right)\left[(295\text{ K})^4 - (306\text{ K})^4\right]$$

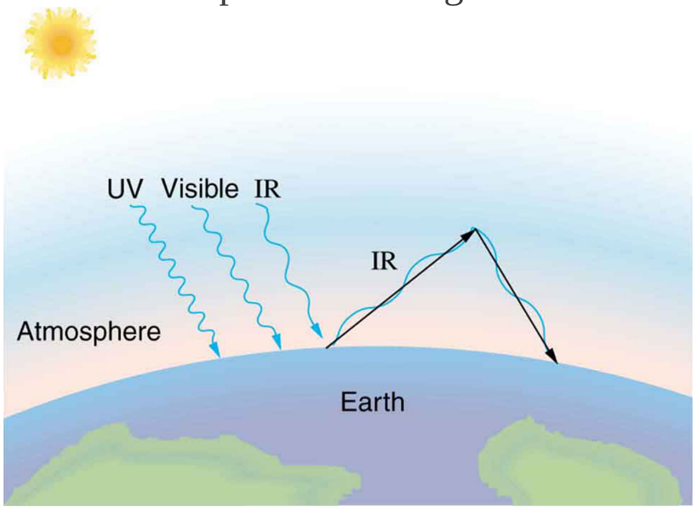
Equation:

$$= -99\text{ J/s} = -99\text{ W}.$$

Discussion

This value is a significant rate of heat transfer to the environment (note the minus sign), considering that a person at rest may produce energy at the rate of 125 W and that conduction and convection will also be transferring energy to the environment. Indeed, we would probably expect this person to feel cold. Clothing significantly reduces heat transfer to the environment by many methods, because clothing slows down both conduction and convection, and has a lower emissivity (especially if it is white) than skin.

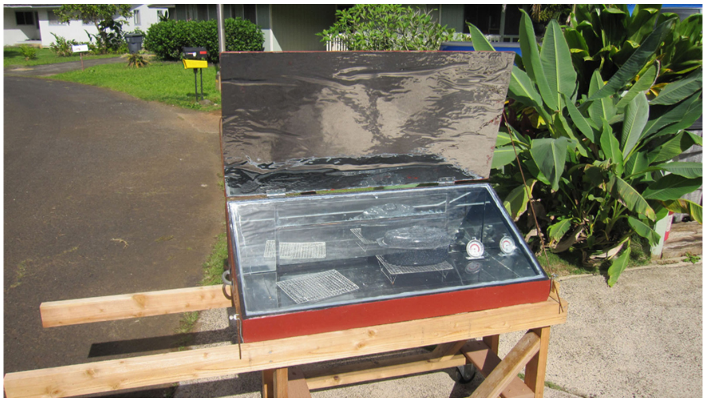
The Earth receives almost all its energy from radiation of the Sun and reflects some of it back into outer space. Because the Sun is hotter than the Earth, the net energy flux is from the Sun to the Earth. However, the rate of energy transfer is less than the equation for the radiative heat transfer would predict because the Sun does not fill the sky. The average emissivity (e) of the Earth is about 0.65, but the calculation of this value is complicated by the fact that the highly reflective cloud coverage varies greatly from day to day. There is a negative feedback (one in which a change produces an effect that opposes that change) between clouds and heat transfer; greater temperatures evaporate more water to form more clouds, which reflect more radiation back into space, reducing the temperature. The often mentioned **greenhouse effect** is directly related to the variation of the Earth's emissivity with radiation type (see the figure given below). The greenhouse effect is a natural phenomenon responsible for providing temperatures suitable for life on Earth. The Earth’s relatively constant temperature is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by carbon dioxide (CO_2) and water (H_2O) in the atmosphere and then re-radiated back to the Earth or into outer space. Re-radiation back to the Earth maintains its surface temperature about 40^oC higher than it would be if there was no atmosphere, similar to the way glass increases temperatures in a greenhouse.



The greenhouse effect is a name given to the trapping of energy in the Earth’s atmosphere by a process similar to that used in greenhouses. The atmosphere, like window glass, is transparent to incoming visible radiation and most of the Sun’s infrared. These wavelengths are absorbed by the Earth and re-emitted as infrared. Since Earth’s temperature is much lower than that of the Sun, the infrared radiated by the Earth has a much longer wavelength. The atmosphere, like glass, traps these longer infrared rays, keeping the Earth warmer than it would otherwise be. The amount of trapping depends on concentrations of trace gases like carbon dioxide, and a change in the concentration of these gases is believed to affect the Earth’s surface temperature.

The greenhouse effect is also central to the discussion of global warming due to emission of carbon dioxide and methane (and other so-called greenhouse gases) into the Earth’s atmosphere from industrial production and farming. Changes in global climate could lead to more intense storms, precipitation changes (affecting agriculture), reduction in rain forest biodiversity, and rising sea levels.

Heating and cooling are often significant contributors to energy use in individual homes. Current research efforts into developing environmentally friendly homes quite often focus on reducing conventional heating and cooling through better building materials, strategically positioning windows to optimize radiation gain from the Sun, and opening spaces to allow convection. It is possible to build a zero-energy house that allows for comfortable living in most parts of the United States with hot and humid summers and cold winters.



This simple but effective solar cooker uses the greenhouse effect and reflective material to trap and retain solar energy. Made of inexpensive, durable materials, it saves money and labor, and is of particular economic value in energy-poor developing countries. (credit: E.B. Kauai)

Conversely, dark space is very cold, about $3\text{K}\ (-\ 454^{\text{o}}\text{F})$, so that the Earth radiates energy into the dark sky. Owing to the fact that clouds have lower emissivity than either oceans or land masses, they reflect some of the radiation back to the surface, greatly reducing heat transfer into dark space, just as they greatly reduce heat transfer into the atmosphere during the day. The rate of heat transfer from soil and grasses can be so rapid that frost may occur on clear summer evenings, even in warm latitudes.

Exercise:

Check Your Understanding

Problem:

What is the change in the rate of the radiated heat by a body at the temperature $T_1 = 20^{\text{o}}\text{C}$ compared to when the body is at the temperature $T_2 = 40^{\text{o}}\text{C}$?

Solution:

The radiated heat is proportional to the fourth power of the *absolute temperature*. Because $T_1 = 293\ \text{K}$ and $T_2 = 313\ \text{K}$, the rate of heat transfer increases by about 30 percent of the original rate.

Note:

Career Connection: Energy Conservation Consultation

The cost of energy is generally believed to remain very high for the foreseeable future. Thus, passive control of heat loss in both commercial and domestic housing will become increasingly important. Energy consultants measure and analyze the flow of energy into and out of houses and ensure that a healthy exchange of air is maintained inside the house. The job prospects for an energy consultant are strong.

Note:

Problem-Solving Strategies for the Methods of Heat Transfer

1. *Examine the situation to determine what type of heat transfer is involved.*
2. *Identify the type(s) of heat transfer—conduction, convection, or radiation.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is very useful.*
4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).*
5. *Solve the appropriate equation for the quantity to be determined (the unknown).*
6. For conduction, equation $\frac{Q}{t}=\frac{\text{kA}\left(T_2-T_1\right)}{d}$ is appropriate. [link] lists thermal conductivities. For convection, determine the amount of matter moved and use equation $Q=\text{mc}\Delta T$, to calculate the heat transfer involved in the temperature change of the fluid. If a phase change accompanies convection, equation $Q=\text{mL}_{\text{f}}$ or $Q=\text{mL}_{\text{v}}$ is appropriate to find the heat transfer involved in the phase change. [link] lists information relevant to phase change. For radiation, equation $\frac{Q_{\text{net}}}{t}=\sigma eA\left(T_2^4-T_1^4\right)$ gives the net heat transfer rate.
7. *Insert the knowns along with their units into the appropriate equation and obtain numerical solutions complete with units.*
8. *Check the answer to see if it is reasonable. Does it make sense?*

Summary

- Radiation is the rate of heat transfer through the emission or absorption of electromagnetic waves.

- The rate of heat transfer depends on the surface area and the fourth power of the absolute temperature:
  Equation:

  $$\frac{Q}{t} = \sigma e A T^4,$$

  where $\sigma = 5.67 \times 10^{-8}\ \text{J/s} \cdot \text{m}^2 \cdot \text{K}^4$ is the Stefan-Boltzmann constant and e is the emissivity of the body. For a black body, $e = 1$ whereas a shiny white or perfect reflector has $e = 0$, with real objects having values of e between 1 and 0. The net rate of heat transfer by radiation is
  Equation:

  $$\frac{Q_{\text{net}}}{t} = \sigma e A \left(T_2^4 - T_1^4\right)$$

  where T_1 is the temperature of an object surrounded by an environment with uniform temperature T_2 and e is the emissivity of the *object*.

Conceptual Questions

Exercise:

Problem:

When watching a daytime circus in a large, dark-colored tent, you sense significant heat transfer from the tent. Explain why this occurs.

Exercise:

Problem:

Satellites designed to observe the radiation from cold (3 K) dark space have sensors that are shaded from the Sun, Earth, and Moon and that are cooled to very low temperatures. Why must the sensors be at low temperature?

Exercise:

Problem: Why are cloudy nights generally warmer than clear ones?

Exercise:

Problem:

Why are thermometers that are used in weather stations shielded from the sunshine? What does a thermometer measure if it is shielded from the sunshine and also if it is not?

Exercise:

Problem:

On average, would Earth be warmer or cooler without the atmosphere? Explain your answer.

Problems & Exercises

Exercise:

Problem:

At what net rate does heat radiate from a 275-m^2 black roof on a night when the roof's temperature is 30.0°C and the surrounding temperature is 15.0°C? The emissivity of the roof is 0.900.

Solution:

-21.7 kW
Note that the negative answer implies heat loss to the surroundings.

Exercise:

Problem:

(a) Cherry-red embers in a fireplace are at 850°C and have an exposed area of $0.200\ \text{m}^2$ and an emissivity of 0.980. The surrounding room has a temperature of 18.0°C. If 50% of the radiant energy enters the room, what is the net rate of radiant heat transfer in kilowatts? (b) Does your answer support the contention that most of the heat transfer into a room by a fireplace comes from infrared radiation?

Exercise:

Problem:

Radiation makes it impossible to stand close to a hot lava flow. Calculate the rate of heat transfer by radiation from $1.00\ \text{m}^2$ of 1200°C fresh lava into 30.0°C surroundings, assuming lava's emissivity is 1.00.

Solution:

$-266\ \text{kW}$

Exercise:

Problem:

(a) Calculate the rate of heat transfer by radiation from a car radiator at $110\,^{\circ}\text{C}$ into a 50.0°C environment, if the radiator has an emissivity of 0.750 and a 1.20-m^2 surface area. (b) Is this a significant fraction of the heat transfer by an automobile engine? To answer this, assume a horsepower of $200\ \text{hp}\ (1.5\ \text{kW})$ and the efficiency of automobile engines as 25%.

Exercise:

Problem:

Find the net rate of heat transfer by radiation from a skier standing in the shade, given the following. She is completely clothed in white (head to foot, including a ski mask), the clothes have an emissivity of 0.200 and a surface temperature of 10.0°C, the surroundings are at -15.0°C, and her surface area is $1.60\ \text{m}^2$.

Solution:

$-36.0\ \text{W}$

Exercise:

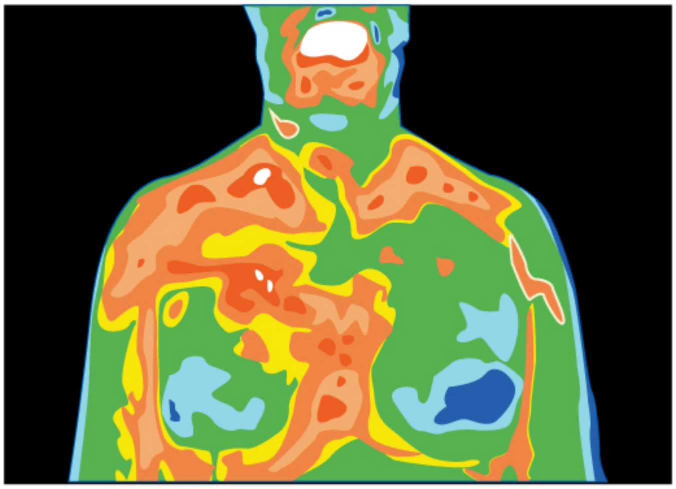
Problem:

Suppose you walk into a sauna that has an ambient temperature of 50.0°C. (a) Calculate the rate of heat transfer to you by radiation given your skin temperature is 37.0°C, the emissivity of skin is 0.98, and the surface area of your body is $1.50\ \text{m}^2$. (b) If all other forms of heat transfer are balanced (the net heat transfer is zero), at what rate will your body temperature increase if your mass is 75.0 kg?

Exercise:

Problem:

Thermography is a technique for measuring radiant heat and detecting variations in surface temperatures that may be medically, environmentally, or militarily meaningful.(a) What is the percent increase in the rate of heat transfer by radiation from a given area at a temperature of 34.0°C compared with that at 33.0°C, such as on a person’s skin? (b) What is the percent increase in the rate of heat transfer by radiation from a given area at a temperature of 34.0°C compared with that at 20.0°C, such as for warm and cool automobile hoods?



Artist’s rendition of a thermograph of a patient’s upper body, showing the distribution of heat represented by different colors.

Solution:

(a) 1.31%

(b) 20.5%

Exercise:

Problem:

The Sun radiates like a perfect black body with an emissivity of exactly 1. (a) Calculate the surface temperature of the Sun, given that it is a sphere with a 7.00×10^{8}-m radius that radiates 3.80×10^{26} W into 3-K space. (b) How much power does the Sun radiate per square meter of its surface? (c) How much power in watts per square meter is that value at the distance of Earth, 1.50×10^{11} m away? (This number is called the solar constant.)

Exercise:

Problem:

A large body of lava from a volcano has stopped flowing and is slowly cooling. The interior of the lava is at 1200^{o}C, its surface is at 450^{o}C, and the surroundings are at 27.0^{o}C. (a) Calculate the rate at which energy is transferred by radiation from $1.00\ \text{m}^2$ of surface lava into the surroundings, assuming the emissivity is 1.00. (b) Suppose heat conduction to the surface occurs at the same rate. What is the thickness of the lava between the 450^{o}C surface and the 1200^{o}C interior, assuming that the lava's conductivity is the same as that of brick?

Solution:

(a) $-15.0\ \text{kW}$

(b) 4.2 cm

Exercise:

Problem:

Calculate the temperature the entire sky would have to be in order to transfer energy by radiation at $1000\ \text{W/m}^2$ —about the rate at which the Sun radiates when it is directly overhead on a clear day. This value is the effective temperature of the sky, a kind of average that takes account of the fact that the Sun occupies only a small part of the sky but is much hotter than the rest. Assume that the body receiving the energy has a temperature of 27.0^{o}C.

Exercise:

Problem:

(a) A shirtless rider under a circus tent feels the heat radiating from the sunlit portion of the tent. Calculate the temperature of the tent canvas based on the following information: The shirtless rider’s skin temperature is 34.0^{o}C and has an emissivity of 0.970. The exposed area of skin is $0.400\ \text{m}^2$. He receives radiation at the rate of 20.0 W—half what you would calculate if the entire region behind him was hot. The rest of the surroundings are at 34.0^{o}C. (b) Discuss how this situation would change if the sunlit side of the tent was nearly pure white and if the rider was covered by a white tunic.

Solution:

(a) 48.5^{o}C

(b) A pure white object reflects more of the radiant energy that hits it, so a white tent would prevent more of the sunlight from heating up the inside of the tent, and the white tunic would prevent that heat which entered the tent from heating the rider. Therefore, with a white tent, the temperature would be lower than 48.5^{o}C, and the rate of radiant heat transferred to the rider would be less than 20.0 W.

Exercise:

Problem: Integrated Concepts

One 30.0^{o}C day the relative humidity is 75.0%, and that evening the temperature drops to 20.0^{o}C, well below the dew point. (a) How many grams of water condense from each cubic meter of air? (b) How much heat transfer occurs by this condensation? (c) What temperature increase could this cause in dry air?

Exercise:

Problem: Integrated Concepts

Large meteors sometimes strike the Earth, converting most of their kinetic energy into thermal energy. (a) What is the kinetic energy of a 10^9 kg meteor moving at 25.0 km/s? (b) If this meteor lands in a deep ocean and 80% of its kinetic energy goes into heating water, how many kilograms of water could it raise by 5.0°C? (c) Discuss how the energy of the meteor is more likely to be deposited in the ocean and the likely effects of that energy.

Solution:

(a) 3×10^{17} J

(b) 1×10^{13} kg

(c) When a large meteor hits the ocean, it causes great tidal waves, dissipating large amount of its energy in the form of kinetic energy of the water.

Exercise:

Problem: **Integrated Concepts**

Frozen waste from airplane toilets has sometimes been accidentally ejected at high altitude. Ordinarily it breaks up and disperses over a large area, but sometimes it holds together and strikes the ground. Calculate the mass of 0°C ice that can be melted by the conversion of kinetic and gravitational potential energy when a $20.0\ \text{kg}$ piece of frozen waste is released at 12.0 km altitude while moving at 250 m/s and strikes the ground at 100 m/s (since less than 20.0 kg melts, a significant mess results).

Exercise:

Problem: **Integrated Concepts**

(a) A large electrical power facility produces 1600 MW of “waste heat,” which is dissipated to the environment in cooling towers by warming air flowing through the towers by 5.00°C. What is the necessary flow rate of air in $\mathrm{m^3/s}$? (b) Is your result consistent with the large cooling towers used by many large electrical power plants?

Solution:

(a) $3.44\times10^5\ \mathrm{m^3/s}$

(b) This is equivalent to 12 million cubic feet of air per second. That is tremendous. This is too large to be dissipated by heating the air by only $5^{\mathrm{o}}\mathrm{C}$. Many of these cooling towers use the circulation of cooler air over warmer water to increase the rate of evaporation. This would allow much smaller amounts of air necessary to remove such a large amount of heat because evaporation removes larger quantities of heat than was considered in part (a).

Exercise:

Problem: Integrated Concepts

(a) Suppose you start a workout on a Stairmaster, producing power at the same rate as climbing 116 stairs per minute. Assuming your mass is 76.0 kg and your efficiency is 20.0%, how long will it take for your body temperature to rise $1.00^{\mathrm{o}}\mathrm{C}$ if all other forms of heat transfer in and out of your body are balanced? (b) Is this consistent with your experience in getting warm while exercising?

Exercise:

Problem: Integrated Concepts

A 76.0-kg person suffering from hypothermia comes indoors and shivers vigorously. How long does it take the heat transfer to increase the person’s body temperature by $2.00^{\mathrm{o}}\mathrm{C}$ if all other forms of heat transfer are balanced?

Solution:

20.9 min

Exercise:

Problem: **Integrated Concepts**

In certain large geographic regions, the underlying rock is hot. Wells can be drilled and water circulated through the rock for heat transfer for the generation of electricity. (a) Calculate the heat transfer that can be extracted by cooling $1.00\ \text{km}^3$ of granite by 100^oC. (b) How long will this take if heat is transferred at a rate of 300 MW, assuming no heat transfers back into the 1.00 km of rock by its surroundings?

Exercise:

Problem: **Integrated Concepts**

Heat transfers from your lungs and breathing passages by evaporating water. (a) Calculate the maximum number of grams of water that can be evaporated when you inhale 1.50 L of 37^oC air with an original relative humidity of 40.0%. (Assume that body temperature is also 37^oC.) (b) How many joules of energy are required to evaporate this amount? (c) What is the rate of heat transfer in watts from this method, if you breathe at a normal resting rate of 10.0 breaths per minute?

Solution:

(a) 3.96×10^{-2} g

(b) 96.2 J

(c) 16.0 W

Exercise:

Problem: **Integrated Concepts**

(a) What is the temperature increase of water falling 55.0 m over Niagara Falls? (b) What fraction must evaporate to keep the temperature constant?

Exercise:

Problem: **Integrated Concepts**

Hot air rises because it has expanded. It then displaces a greater volume of cold air, which increases the buoyant force on it. (a) Calculate the ratio of the buoyant force to the weight of 50.0°C air surrounded by 20.0°C air. (b) What energy is needed to cause $1.00\ \text{m}^3$ of air to go from 20.0°C to 50.0°C? (c) What gravitational potential energy is gained by this volume of air if it rises 1.00 m? Will this cause a significant cooling of the air?

Solution:

(a) 1.102

(b) $2.79 \times 10^4\ \text{J}$

(c) 12.6 J. This will not cause a significant cooling of the air because it is much less than the energy found in part (b), which is the energy required to warm the air from 20.0°C to 50.0°C.

Exercise:

Problem: **Unreasonable Results**

(a) What is the temperature increase of an 80.0 kg person who consumes 2500 kcal of food in one day with 95.0% of the energy transferred as heat to the body? (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?

Solution:

(a) 36°C

(b) Any temperature increase greater than about 3°C would be unreasonably large. In this case the final temperature of the person would rise to $73^{\circ}\text{C}\ (163^{\circ}\text{F})$.

(c) The assumption of 95% heat retention is unreasonable.

Exercise:

Problem: Unreasonable Results

A slightly deranged Arctic inventor surrounded by ice thinks it would be much less mechanically complex to cool a car engine by melting ice on it than by having a water-cooled system with a radiator, water pump, antifreeze, and so on. (a) If 80.0% of the energy in 1.00 gal of gasoline is converted into “waste heat” in a car engine, how many kilograms of $0^{\circ}\mathrm{C}$ ice could it melt? (b) Is this a reasonable amount of ice to carry around to cool the engine for 1.00 gal of gasoline consumption? (c) What premises or assumptions are unreasonable?

Exercise:

Problem: Unreasonable Results

(a) Calculate the rate of heat transfer by conduction through a window with an area of $1.00\ \mathrm{m}^2$ that is 0.750 cm thick, if its inner surface is at $22.0^{\circ}\mathrm{C}$ and its outer surface is at $35.0^{\circ}\mathrm{C}$. (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?

Solution:

(a) 1.46 kW

(b) Very high power loss through a window. An electric heater of this power can keep an entire room warm.

(c) The surface temperatures of the window do not differ by as great an amount as assumed. The inner surface will be warmer, and the outer surface will be cooler.

Exercise:

Problem: Unreasonable Results

A meteorite 1.20 cm in diameter is so hot immediately after penetrating the atmosphere that it radiates 20.0 kW of power. (a) What is its temperature, if the surroundings are at 20.0ºC and it has an emissivity of 0.800? (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?

Exercise:

Problem: Construct Your Own Problem

Consider a new model of commercial airplane having its brakes tested as a part of the initial flight permission procedure. The airplane is brought to takeoff speed and then stopped with the brakes alone. Construct a problem in which you calculate the temperature increase of the brakes during this process. You may assume most of the kinetic energy of the airplane is converted to thermal energy in the brakes and surrounding materials, and that little escapes. Note that the brakes are expected to become so hot in this procedure that they ignite and, in order to pass the test, the airplane must be able to withstand the fire for some time without a general conflagration.

Exercise:

Problem: Construct Your Own Problem

Consider a person outdoors on a cold night. Construct a problem in which you calculate the rate of heat transfer from the person by all three heat transfer methods. Make the initial circumstances such that at rest the person will have a net heat transfer and then decide how much physical activity of a chosen type is necessary to balance the rate of heat transfer. Among the things to consider are the size of the person, type of clothing, initial metabolic rate, sky conditions, amount of water evaporated, and volume of air breathed. Of course, there are many other factors to consider and your instructor may wish to guide you in the assumptions made as well as the detail of analysis and method of presenting your results.

Glossary

emissivity

measure of how well an object radiates

greenhouse effect

warming of the Earth that is due to gases such as carbon dioxide and methane that absorb infrared radiation from the Earth’s surface and reradiate it in all directions, thus sending a fraction of it back toward the surface of the Earth

net rate of heat transfer by radiation

is $\frac{Q_{\text{net}}}{t} = \sigma eA\left(T_2^4 - T_1^4\right)$

radiation

energy transferred by electromagnetic waves directly as a result of a temperature difference

Stefan-Boltzmann law of radiation

$\frac{Q}{t} = \sigma eAT^4$, where σ is the Stefan-Boltzmann constant, A is the surface area of the object, T is the absolute temperature, and e is the emissivity

Introduction to Thermodynamics class="introduction"

A steam engine uses heat transfer to do work. Tourists regularly ride this narrow-gauge steam engine train near the San Juan Skyway in Durango, Colorado, part of the National Scenic Byways Program. (credit: Dennis Adams)

Heat transfer is energy in transit, and it can be used to do work. It can also be converted to any other form of energy. A car engine, for example, burns fuel for heat transfer into a gas. Work is done by the gas as it exerts a force through a distance, converting its energy into a variety of other forms—into the car's kinetic or gravitational potential energy; into electrical energy to run the spark plugs, radio, and lights; and back into stored energy in the car's battery. But most of the heat transfer produced from burning fuel in the engine does not do work on the gas. Rather, the energy is released into the environment, implying that the engine is quite inefficient.

It is often said that modern gasoline engines cannot be made to be significantly more efficient. We hear the same about heat transfer to electrical energy in large power stations, whether they are coal, oil, natural gas, or nuclear powered. Why is that the case? Is the inefficiency caused by design problems that could be solved with better engineering and superior materials? Is it part of some money-making conspiracy by those who sell energy? Actually, the truth is more interesting, and reveals much about the nature of heat transfer.

Basic physical laws govern how heat transfer for doing work takes place and place insurmountable limits onto its efficiency. This chapter will explore these laws as well as many applications and concepts associated with them. These topics are part of *thermodynamics*—the study of heat transfer and its relationship to doing work.

The First Law of Thermodynamics

- Define the first law of thermodynamics.
- Describe how conservation of energy relates to the first law of thermodynamics.
- Identify instances of the first law of thermodynamics working in everyday situations, including biological metabolism.
- Calculate changes in the internal energy of a system, after accounting for heat transfer and work done.

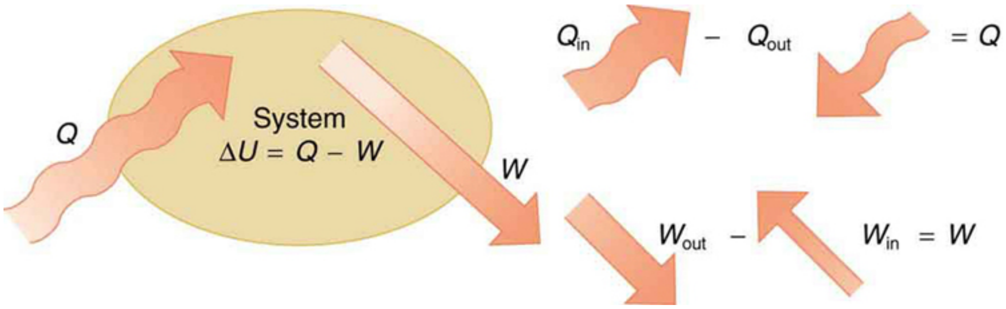
This boiling tea kettle represents energy in motion. The water in the kettle is turning to water vapor because heat is being transferred from the stove to the kettle. As the entire system gets hotter, work is done—from the evaporation of the water to the whistling of the kettle. (credit: Gina Hamilton)

If we are interested in how heat transfer is converted into doing work, then the conservation of energy principle is important. The first law of thermodynamics applies the conservation of energy principle to systems where heat transfer and doing work are the methods of transferring energy into and out of the system. The **first law of thermodynamics** states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system. In equation form, the first law of thermodynamics is

Equation:

$$\Delta U = Q - W.$$

Here ΔU is the *change in internal energy* U of the system. Q is the *net heat transferred into the system*—that is, Q is the sum of all heat transfer into and out of the system. W is the *net work done by the system*—that is, W is the sum of all work done on or by the system. We use the following sign conventions: if Q is positive, then there is a net heat transfer into the system; if W is positive, then there is net work done by the system. So positive Q adds energy to the system and positive W takes energy from the system. Thus $\Delta U = Q - W$. Note also that if more heat transfer into the system occurs than work done, the difference is stored as internal energy. Heat engines are a good example of this—heat transfer into them takes place so that they can do work. (See [link].) We will now examine Q, W, and ΔU further.



The first law of thermodynamics is the conservation-of-energy principle stated for a system where heat and work are the methods of transferring energy for a system in thermal equilibrium. Q represents the net heat transfer—it is the sum of all heat transfers into and out of the system. Q is positive for net heat transfer *into* the system. W is the total work done on and by the system. W is positive when more work is done *by* the system than on it. The change in the internal energy of the system, ΔU, is related to heat and work by the first law of thermodynamics, $\Delta U = Q - W$.

Note:
Making Connections: Law of Thermodynamics and Law of Conservation of Energy
The first law of thermodynamics is actually the law of conservation of energy stated in a form most useful in thermodynamics. The first law gives the relationship between heat transfer, work done, and the change in internal energy of a system.

Heat *Q* and Work *W*

Heat transfer (Q) and doing work (W) are the two everyday means of bringing energy into or taking energy out of a system. The processes are quite different. Heat transfer, a less organized process, is driven by temperature differences. Work, a quite organized process, involves a macroscopic force exerted through a distance. Nevertheless, heat and work can produce identical results.For example, both can cause a temperature increase. Heat transfer into a system, such as when the Sun warms the air in a bicycle tire, can increase its temperature, and so can work done on the system, as when the bicyclist pumps air into the tire. Once the temperature increase has occurred, it is impossible to tell whether it was caused by heat transfer or by doing work. This uncertainty is an important point. Heat transfer and work are both energy in transit—neither is stored as such in a system. However, both can change the internal energy U of a system. Internal energy is a form of energy completely different from either heat or work.

Internal Energy *U*

We can think about the internal energy of a system in two different but consistent ways. The first is the atomic and molecular view, which examines the system on the atomic and molecular scale. The **internal energy** U of a system is the sum of the kinetic and potential energies of its atoms and molecules. Recall that kinetic plus potential energy is called mechanical energy. Thus internal energy is the sum of atomic and molecular mechanical energy. Because it is impossible to keep track of all individual atoms and molecules, we must deal with averages and distributions. A second way to view the internal energy of a system is in terms of its macroscopic characteristics, which are very similar to atomic and molecular average values.

Macroscopically, we define the change in internal energy ΔU to be that given by the first law of thermodynamics:

Equation:

$$\Delta U = Q - W.$$

Many detailed experiments have verified that $\Delta U = Q - W$, where ΔU is the change in total kinetic and potential energy of all atoms and molecules in a system. It has also been determined experimentally that the internal energy U of a system depends only on the state of the system and *not how it reached that state*. More specifically, U is found to be a function of a few macroscopic quantities (pressure, volume, and temperature, for example), independent of past history such as whether there has been heat transfer or work done. This independence means that if we know the state of a system, we can calculate changes in its internal energy U from a few macroscopic variables.

Note:

Making Connections: Macroscopic and Microscopic

In thermodynamics, we often use the macroscopic picture when making calculations of how a system behaves, while the atomic and molecular picture gives underlying explanations in terms of averages and distributions. We shall see this again in later sections of this chapter. For example, in the topic of entropy, calculations will be made using the atomic and molecular view.

To get a better idea of how to think about the internal energy of a system, let us examine a system going from State 1 to State 2. The system has internal energy U_1 in State 1, and it has internal energy U_2 in State 2, no matter how it got to either state. So the change in internal energy $\Delta U = U_2 - U_1$ is independent of what caused the change. In other words, ΔU *is independent of path*. By path, we mean the method of getting from the starting point to the ending point. Why is this independence important? Note that $\Delta U = Q - W$. Both Q and W *depend on path*, but ΔU does not. This path independence means that internal energy U is easier to consider than either heat transfer or work done.

Example:

Calculating Change in Internal Energy: The Same Change in U is Produced by Two Different Processes

(a) Suppose there is heat transfer of 40.00 J to a system, while the system does 10.00 J of work. Later, there is heat transfer of 25.00 J out of the system while 4.00 J of work is done on the system. What is the net change in internal energy of the system?

(b) What is the change in internal energy of a system when a total of 150.00 J of heat transfer occurs out of (from) the system and 159.00 J of work is done on the system? (See [link]).

Strategy

In part (a), we must first find the net heat transfer and net work done from the given information. Then the first law of thermodynamics

$(\Delta U = Q - W)$ can be used to find the change in internal energy. In part (b), the net heat transfer and work done are given, so the equation can be used directly.

Solution for (a)

The net heat transfer is the heat transfer into the system minus the heat transfer out of the system, or

Equation:

$$Q = 40.00 \text{ J} - 25.00 \text{ J} = 15.00 \text{ J}.$$

Similarly, the total work is the work done by the system minus the work done on the system, or

Equation:

$$W = 10.00 \text{ J} - 4.00 \text{ J} = 6.00 \text{ J}.$$

Thus the change in internal energy is given by the first law of thermodynamics:

Equation:

$$\Delta U = Q - W = 15.00 \text{ J} - 6.00 \text{ J} = 9.00 \text{ J}.$$

We can also find the change in internal energy for each of the two steps. First, consider 40.00 J of heat transfer in and 10.00 J of work out, or

Equation:

$$\Delta U_1 = Q_1 - W_1 = 40.00 \text{ J} - 10.00 \text{ J} = 30.00 \text{ J}.$$

Now consider 25.00 J of heat transfer out and 4.00 J of work in, or

Equation:

$$\Delta U_2 = Q_2 - W_2 = - 25.00 \text{ J} - (-4.00 \text{ J}) = -21.00 \text{ J}.$$

The total change is the sum of these two steps, or

Equation:

$$\Delta U = \Delta U_1 + \Delta U_2 = 30.00 \text{ J} + (-21.00 \text{ J}) = 9.00 \text{ J}.$$

Discussion on (a)

No matter whether you look at the overall process or break it into steps, the change in internal energy is the same.

Solution for (b)

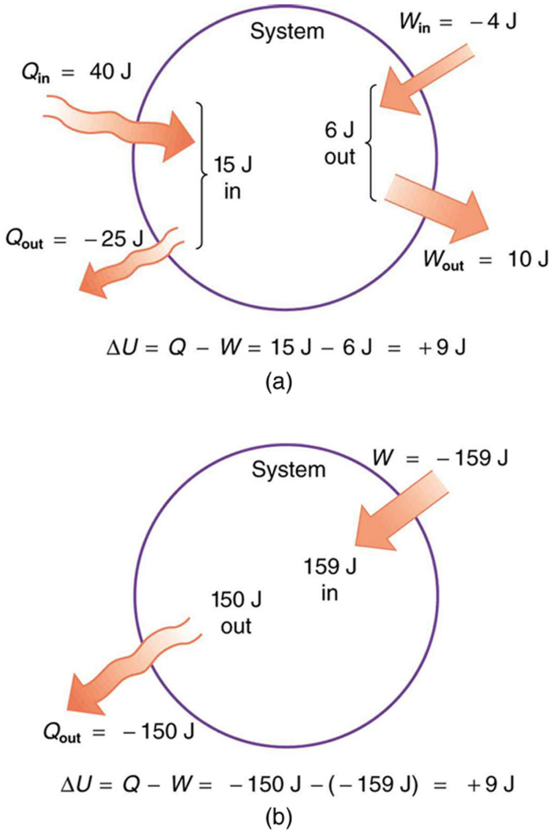
Here the net heat transfer and total work are given directly to be $Q = -150.00\text{ J}$ and $W = -159.00\text{ J}$, so that

Equation:

$$\Delta U = Q - W = -150.00\text{ J} - (-159.00\text{ J}) = 9.00\text{ J}.$$

Discussion on (b)

A very different process in part (b) produces the same 9.00-J change in internal energy as in part (a). Note that the change in the system in both parts is related to ΔU and not to the individual Qs or Ws involved. The system ends up in the *same* state in both (a) and (b). Parts (a) and (b) present two different paths for the system to follow between the same starting and ending points, and the change in internal energy for each is the same—it is independent of path.



Two different processes produce the same change in a system. (a) A total of 15.00 J of heat transfer occurs into the system, while work takes out a total of 6.00 J. The change in internal energy is $\Delta U = Q - W = 9.00\ \text{J}$. (b) Heat transfer removes 150.00 J from the system while work puts 159.00 J into it, producing an increase of 9.00 J in internal energy. If the system starts out in the same state in (a) and (b), it will end up in the same final state in either case—its final state is related to internal

energy, not how that energy
was acquired.

Human Metabolism and the First Law of Thermodynamics

Human metabolism is the conversion of food into heat transfer, work, and stored fat. Metabolism is an interesting example of the first law of thermodynamics in action. We now take another look at these topics via the first law of thermodynamics. Considering the body as the system of interest, we can use the first law to examine heat transfer, doing work, and internal energy in activities ranging from sleep to heavy exercise. What are some of the major characteristics of heat transfer, doing work, and energy in the body? For one, body temperature is normally kept constant by heat transfer to the surroundings. This means Q is negative. Another fact is that the body usually does work on the outside world. This means W is positive. In such situations, then, the body loses internal energy, since $\Delta U = Q - W$ is negative.

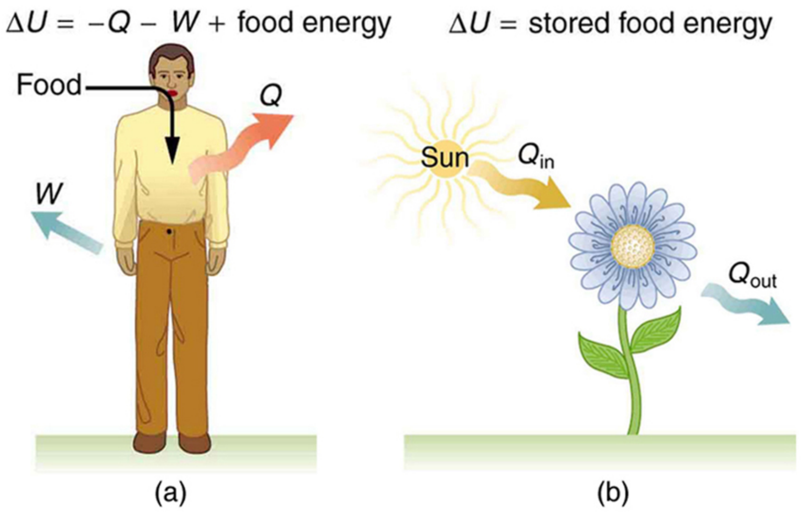
Now consider the effects of eating. Eating increases the internal energy of the body by adding chemical potential energy (this is an unromantic view of a good steak). The body *metabolizes* all the food we consume. Basically, metabolism is an oxidation process in which the chemical potential energy of food is released. This implies that food input is in the form of work. Food energy is reported in a special unit, known as the Calorie. This energy is measured by burning food in a calorimeter, which is how the units are determined.

In chemistry and biochemistry, one calorie (spelled with a *lowercase* c) is defined as the energy (or heat transfer) required to raise the temperature of one gram of pure water by one degree Celsius. Nutritionists and weight-watchers tend to use the *dietary* calorie, which is frequently called a Calorie (spelled with a *capital* C). One food Calorie is the energy needed to raise the temperature of one *kilogram* of water by one degree Celsius. This means that one dietary Calorie is equal to one kilocalorie for the chemist, and one must be careful to avoid confusion between the two.

Again, consider the internal energy the body has lost. There are three places this internal energy can go—to heat transfer, to doing work, and to stored fat (a tiny fraction also goes to cell repair and growth). Heat transfer and doing work take internal energy out of the body, and food puts it back. If you eat just the right amount of food, then your average internal energy remains constant. Whatever you lose to heat transfer and doing work is replaced by food, so that, in the long run, $\Delta U = 0$. If you overeat repeatedly, then ΔU is always positive, and your body stores this extra internal energy as fat. The reverse is true if you eat too little. If ΔU is negative for a few days, then the body metabolizes its own fat to maintain body temperature and do work that takes energy from the body. This process is how dieting produces weight loss.

Life is not always this simple, as any dieter knows. The body stores fat or metabolizes it only if energy intake changes for a period of several days. Once you have been on a major diet, the next one is less successful because your body alters the way it responds to low energy intake. Your basal metabolic rate (BMR) is the rate at which food is converted into heat transfer and work done while the body is at complete rest. The body adjusts its basal metabolic rate to partially compensate for over-eating or under-eating. The body will decrease the metabolic rate rather than eliminate its own fat to replace lost food intake. You will chill more easily and feel less energetic as a result of the lower metabolic rate, and you will not lose weight as fast as before. Exercise helps to lose weight, because it produces both heat transfer from your body and work, and raises your metabolic rate even when you are at rest. Weight loss is also aided by the quite low efficiency of the body in converting internal energy to work, so that the loss of internal energy resulting from doing work is much greater than the work done.It should be noted, however, that living systems are not in thermalequilibrium.

The body provides us with an excellent indication that many thermodynamic processes are *irreversible*. An irreversible process can go in one direction but not the reverse, under a given set of conditions. For example, although body fat can be converted to do work and produce heat transfer, work done on the body and heat transfer into it cannot be converted to body fat. Otherwise, we could skip lunch by sunning ourselves or by walking down stairs. Another example of an irreversible thermodynamic process is photosynthesis. This process is the intake of one form of energy—light—by plants and its conversion to chemical potential energy. Both applications of the first law of thermodynamics are illustrated in [link]. One great advantage of conservation laws such as the first law of thermodynamics is that they accurately describe the beginning and ending points of complex processes, such as metabolism and photosynthesis, without regard to the complications in between. [link] presents a summary of terms relevant to the first law of thermodynamics.


$\Delta U = -Q - W +$ food energy
$\Delta U =$ stored food energy
Food
Q
W
Sun
Q_{in}
Q_{out}
(a)
(b)

(a) The first law of thermodynamics applied to metabolism. Heat transferred out of the body (Q) and work done by the body (W) remove internal energy, while food intake replaces it. (Food intake may be considered as work done on the body.) (b) Plants convert part of the radiant heat transfer in sunlight to stored chemical energy, a process called photosynthesis.

Term	Definition
U	Internal energy—the sum of the kinetic and potential energies of a system’s atoms and molecules. Can be divided into many subcategories, such as thermal and chemical energy. Depends only on the state of a system (such as its P, V, and T), not on how the energy entered the system. Change in internal energy is path independent.
Q	Heat—energy transferred because of a temperature difference. Characterized by random molecular motion. Highly dependent on path. Q entering a system is positive.
W	Work—energy transferred by a force moving through a distance. An organized, orderly process. Path dependent. W done by a system (either against an external force or to increase the volume of the system) is positive.

Summary of Terms for the First Law of Thermodynamics, $\Delta U = Q - W$

Section Summary

- The first law of thermodynamics is given as $\Delta U = Q - W$, where ΔU is the change in internal energy of a system, Q is the net heat transfer (the sum of all heat transfer into and out of the system), and W is the net work done (the sum of all work done on or by the system).
- Both Q and W are energy in transit; only ΔU represents an independent quantity capable of being stored.

- The internal energy U of a system depends only on the state of the system and not how it reached that state.
- Metabolism of living organisms, and photosynthesis of plants, are specialized types of heat transfer, doing work, and internal energy of systems.

Conceptual Questions

Exercise:

Problem:

Describe the photo of the tea kettle at the beginning of this section in terms of heat transfer, work done, and internal energy. How is heat being transferred? What is the work done and what is doing it? How does the kettle maintain its internal energy?

Exercise:

Problem:

The first law of thermodynamics and the conservation of energy, as discussed in Conservation of Energy, are clearly related. How do they differ in the types of energy considered?

Exercise:

Problem:

Heat transfer Q and work done W are always energy in transit, whereas internal energy U is energy stored in a system. Give an example of each type of energy, and state specifically how it is either in transit or resides in a system.

Exercise:

Problem:

How do heat transfer and internal energy differ? In particular, which can be stored as such in a system and which cannot?

Exercise:

Problem:

If you run down some stairs and stop, what happens to your kinetic energy and your initial gravitational potential energy?

Exercise:

Problem:

Give an explanation of how food energy (calories) can be viewed as molecular potential energy (consistent with the atomic and molecular definition of internal energy).

Exercise:

Problem:

Identify the type of energy transferred to your body in each of the following as either internal energy, heat transfer, or doing work: (a) basking in sunlight; (b) eating food; (c) riding an elevator to a higher floor.

Problems & Exercises

Exercise:

Problem:

What is the change in internal energy of a car if you put 12.0 gal of gasoline into its tank? The energy content of gasoline is $1.3 \times 10^8\ \text{J/gal}$. All other factors, such as the car's temperature, are constant.

Solution:

$1.6 \times 10^9\ \text{J}$

Exercise:

Problem:

How much heat transfer occurs from a system, if its internal energy decreased by 150 J while it was doing 30.0 J of work?

Exercise:

Problem:

A system does $1.80\times10^{8}\text{ J}$ of work while $7.50\times10^{8}\text{ J}$ of heat transfer occurs to the environment. What is the change in internal energy of the system assuming no other changes (such as in temperature or by the addition of fuel)?

Solution:

$-9.30\times10^{8}\text{ J}$

Exercise:

Problem:

What is the change in internal energy of a system which does $4.50\times10^{5}\text{ J}$ of work while $3.00\times10^{6}\text{ J}$ of heat transfer occurs into the system, and $8.00\times10^{6}\text{ J}$ of heat transfer occurs to the environment?

Exercise:

Problem:

Suppose a woman does 500 J of work and 9500 J of heat transfer occurs into the environment in the process. (a) What is the decrease in her internal energy, assuming no change in temperature or consumption of food? (That is, there is no other energy transfer.) (b) What is her efficiency?

Solution:

(a) $-1.0\times10^{4}\text{ J}$, or -2.39 kcal

(b) 5.00%

Exercise:

Problem:

(a) How much food energy will a man metabolize in the process of doing 35.0 kJ of work with an efficiency of 5.00%? (b) How much heat transfer occurs to the environment to keep his temperature constant? Explicitly show how you follow the steps in the Problem-Solving Strategy for thermodynamics found in Problem-Solving Strategies for Thermodynamics.

Exercise:

Problem:

(a) What is the average metabolic rate in watts of a man who metabolizes 10,500 kJ of food energy in one day? (b) What is the maximum amount of work in joules he can do without breaking down fat, assuming a maximum efficiency of 20.0%? (c) Compare his work output with the daily output of a 187-W (0.250-horsepower) motor.

Solution:

(a) 122 W

(b) 2.10×10^{6} J

(c) Work done by the motor is 1.61×10^{7} J ;thus the motor produces 7.67 times the work done by the man

Exercise:

Problem:

(a) How long will the energy in a 1470-kJ (350-kcal) cup of yogurt last in a woman doing work at the rate of 150 W with an efficiency of 20.0% (such as in leisurely climbing stairs)? (b) Does the time found in part (a) imply that it is easy to consume more food energy than you can reasonably expect to work off with exercise?

Exercise:

Problem:

(a) A woman climbing the Washington Monument metabolizes 6.00×10^2 kJ of food energy. If her efficiency is 18.0%, how much heat transfer occurs to the environment to keep her temperature constant? (b) Discuss the amount of heat transfer found in (a). Is it consistent with the fact that you quickly warm up when exercising?

Solution:

(a) 492 kJ

(b) This amount of heat is consistent with the fact that you warm quickly when exercising. Since the body is inefficient, the excess heat produced must be dissipated through sweating, breathing, etc.

Glossary

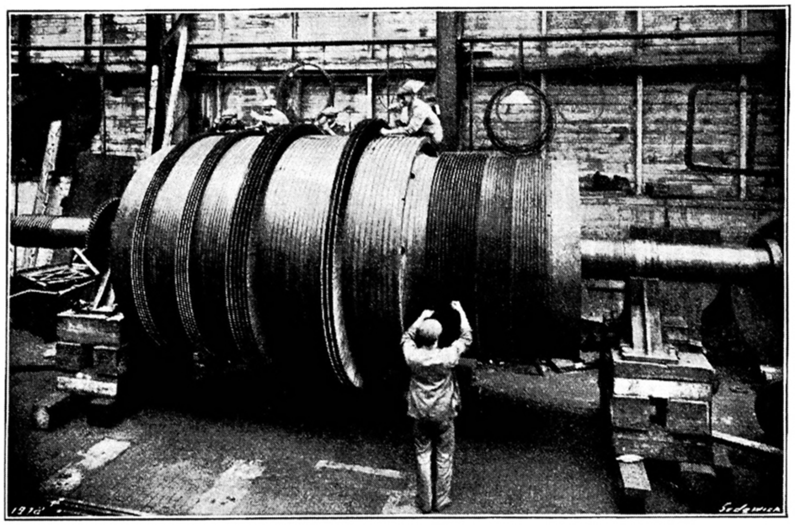
first law of thermodynamics
: states that the change in internal energy of a system equals the net heat transfer *into* the system minus the net work done *by* the system

internal energy
: the sum of the kinetic and potential energies of a system's atoms and molecules

human metabolism
: conversion of food into heat transfer, work, and stored fat

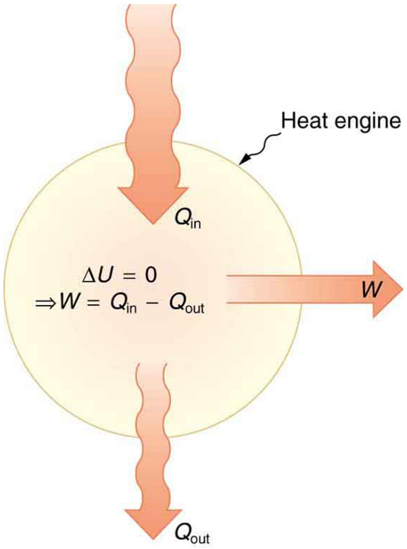
The First Law of Thermodynamics and Some Simple Processes

- Describe the processes of a simple heat engine.
- Explain the differences among the simple thermodynamic processes—isobaric, isochoric, isothermal, and adiabatic.
- Calculate total work done in a cyclical thermodynamic process.

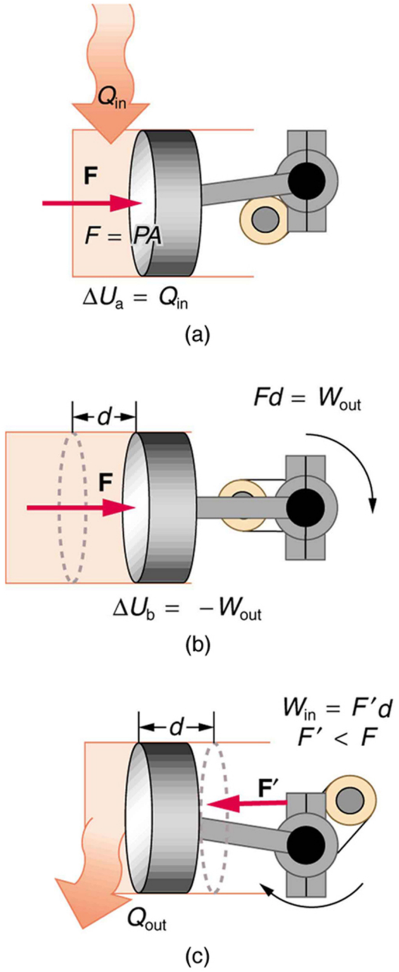


Beginning with the Industrial Revolution, humans have harnessed power through the use of the first law of thermodynamics, before we even understood it completely. This photo, of a steam engine at the Turbinia Works, dates from 1911, a mere 61 years after the first explicit statement of the first law of thermodynamics by Rudolph Clausius. (credit: public domain; author unknown)

One of the most important things we can do with heat transfer is to use it to do work for us. Such a device is called a **heat engine**. Car engines and steam turbines that generate electricity are examples of heat engines. [link] shows schematically how the first law of thermodynamics applies to the typical heat engine.



Schematic representation of a heat engine, governed, of course, by the first law of thermodynamics. It is impossible to devise a system where $Q_{\text{out}} = 0$, that is, in which no heat transfer occurs to the environment.



(a) Heat transfer to the gas in a cylinder increases the internal energy of the gas, creating higher pressure and temperature. (b) The force exerted on the movable cylinder does work as the gas expands. Gas pressure and temperature decrease when it expands, indicating that the gas’s internal energy has been decreased by doing work. (c) Heat transfer to the environment further reduces pressure in the gas so that the piston can be more easily returned to its starting position.

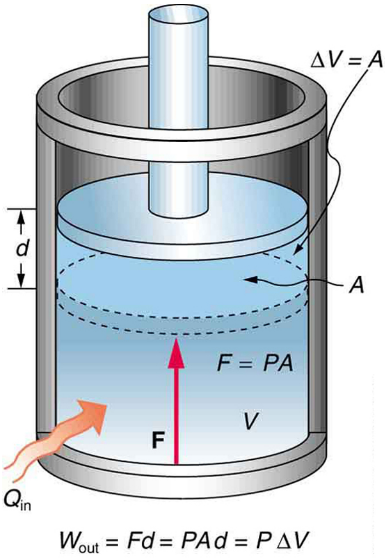
The illustrations above show one of the ways in which heat transfer does work. Fuel combustion produces heat transfer to a gas in a cylinder, increasing the pressure of the gas and thereby the force it exerts on a movable piston. The gas does work on the outside world, as this force moves the piston through some distance. Heat transfer to the gas cylinder results in work being done. To repeat this process, the piston needs to be returned to its starting point. Heat transfer now occurs from the gas to the surroundings so that its pressure decreases, and a force is exerted by the surroundings to push the piston back through some distance. Variations of this process are employed daily in hundreds of millions of heat engines. We will examine heat engines in detail in the next section. In this section, we consider some of the simpler underlying processes on which heat engines are based.

PV Diagrams and their Relationship to Work Done on or by a Gas

A process by which a gas does work on a piston at constant pressure is called an **isobaric process**. Since the pressure is constant, the force exerted is constant and the work done is given as

Equation:

$$P\Delta V.$$



An isobaric expansion of a gas requires heat transfer to keep the pressure constant. Since pressure is constant, the work done is $P\Delta V$.

Equation:

$$W = \text{Fd}$$

See the symbols as shown in [link]. Now $F = \text{PA}$, and so
Equation:

$$W = \text{PAd}.$$

Because the volume of a cylinder is its cross-sectional area A times its length d, we see that $\mathrm{Ad} = \Delta V$, the change in volume; thus,

Equation:

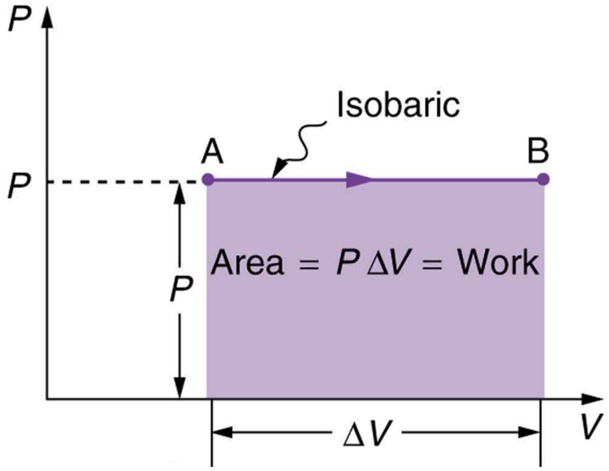
$$W = P\Delta V \text{ (isobaric process).}$$

Note that if ΔV is positive, then W is positive, meaning that work is done *by* the gas on the outside world.

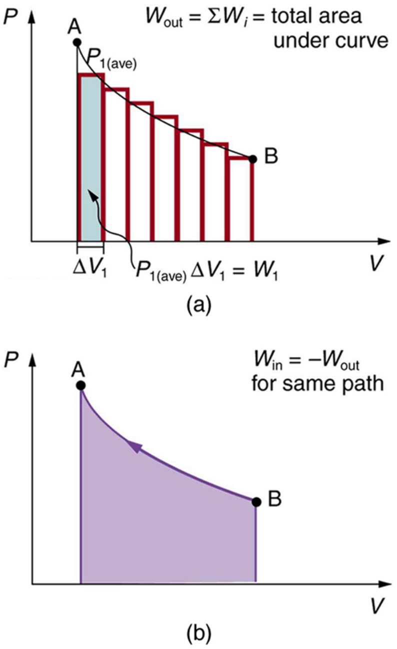
(Note that the pressure involved in this work that we've called P is the pressure of the gas *inside* the tank. If we call the pressure outside the tank P_{ext}, an expanding gas would be working *against* the external pressure; the work done would therefore be $W = -P_{\text{ext}}\Delta V$ (isobaric process). Many texts use this definition of work, and not the definition based on internal pressure, as the basis of the First Law of Thermodynamics. This definition reverses the sign conventions for work, and results in a statement of the first law that becomes $\Delta U = Q + W$.)

It is not surprising that $W = P\Delta V$, since we have already noted in our treatment of fluids that pressure is a type of potential energy per unit volume and that pressure in fact has units of energy divided by volume. We also noted in our discussion of the ideal gas law that PV has units of energy. In this case, some of the energy associated with pressure becomes work.

[link] shows a graph of pressure versus volume (that is, a PV diagram for an isobaric process. You can see in the figure that the work done is the area under the graph. This property of PV diagrams is very useful and broadly applicable: *the work done on or by a system in going from one state to another equals the area under the curve on a* PV *diagram*.



A graph of pressure versus volume for a constant-pressure, or isobaric, process, such as the one shown in [link]. The area under the curve equals the work done by the gas, since $W = P\Delta V$.



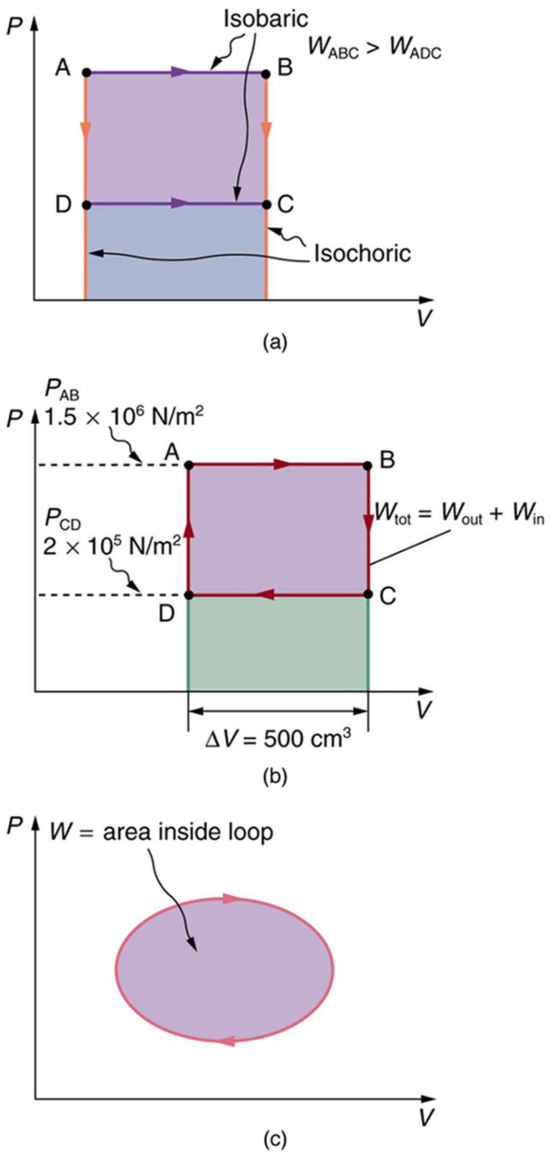
(a)

(b)

(a) A PV diagram in which pressure varies as well as volume. The work done for each interval is its average pressure times the change in volume, or the area under the curve over that interval. Thus the total area under the curve equals the total work done. (b) Work must be done on the system to follow the reverse path. This is interpreted as a negative area under the curve.

We can see where this leads by considering [link](a), which shows a more general process in which both pressure and volume change. The area under the curve is closely approximated by dividing it into strips, each having an average constant pressure $P_{i(\text{ave})}$. The work done is $W_i = P_{i(\text{ave})}\Delta V_i$ for each strip, and the total work done is the sum of the W_i. Thus the total work done is the total area under the curve. If the path is reversed, as in [link](b), then work is done on the system. The area under the curve in that case is negative, because ΔV is negative.

PV diagrams clearly illustrate that *the work done depends on the path taken and not just the endpoints*. This path dependence is seen in [link](a), where more work is done in going from A to C by the path via point B than by the path via point D. The vertical paths, where volume is constant, are called **isochoric** processes. Since volume is constant, $\Delta V = 0$, and no work is done in an isochoric process. Now, if the system follows the cyclical path ABCDA, as in [link](b), then the total work done is the area inside the loop. The negative area below path CD subtracts, leaving only the area inside the rectangle. In fact, the work done in any cyclical process (one that returns to its starting point) is the area inside the loop it forms on a PV diagram, as [link](c) illustrates for a general cyclical process. Note that the loop must be traversed in the clockwise direction for work to be positive—that is, for there to be a net work output.



(a) The work done in going from A to C depends on path. The work is greater for the path ABC than for the path ADC, because the former is at higher pressure. In both cases, the work done is the area under the path. This area is greater for path ABC. (b) The total work done in the cyclical process

ABCDA is the area inside the loop, since the negative area below CD subtracts out, leaving just the area inside the rectangle. (The values given for the pressures and the change in volume are intended for use in the example below.) (c) The area inside any closed loop is the work done in the cyclical process. If the loop is traversed in a clockwise direction, W is positive—it is work done on the outside environment. If the loop is traveled in a counter-clockwise direction, W is negative—it is work that is done to the system.

Example:

Total Work Done in a Cyclical Process Equals the Area Inside the Closed Loop on a *PV* Diagram

Calculate the total work done in the cyclical process ABCDA shown in [link](b) by the following two methods to verify that work equals the area inside the closed loop on the PV diagram. (Take the data in the figure to be precise to three significant figures.) (a) Calculate the work done along each segment of the path and add these values to get the total work. (b) Calculate the area inside the rectangle ABCDA.

Strategy

To find the work along any path on a PV diagram, you use the fact that work is pressure times change in volume, or $W = P\Delta V$. So in part (a), this value is calculated for each leg of the path around the closed loop.

Solution for (a)

The work along path AB is

Equation:

$$\begin{aligned} W_{\text{AB}} &= P_{\text{AB}}\Delta V_{\text{AB}} \\ &= (1.50\times10^{6}\ \text{N/m}^{2})(5.00\times10^{–4}\ \text{m}^{3}) = 750\ \text{J}. \end{aligned}$$

Since the path BC is isochoric, $\Delta V_{\text{BC}} = 0$, and so $W_{\text{BC}} = 0$. The work along path CD is negative, since ΔV_{CD} is negative (the volume decreases). The work is

Equation:

$$\begin{aligned} W_{\text{CD}} &= P_{\text{CD}}\Delta V_{\text{CD}} \\ &= (2.00\times10^{5}\ \text{N/m}^{2})(–5.00\times10^{–4}\ \text{m}^{3}) = –100\ \text{J}. \end{aligned}$$

Again, since the path DA is isochoric, $\Delta V_{\text{DA}} = 0$, and so $W_{\text{DA}} = 0$. Now the total work is

Equation:

$$\begin{aligned} W &= W_{\text{AB}} + W_{\text{BC}} + W_{\text{CD}} + W_{\text{DA}} \\ &= 750\ \text{J} + 0 + (-100\text{J}) + 0 = 650\ \text{J}. \end{aligned}$$

Solution for (b)

The area inside the rectangle is its height times its width, or

Equation:

$$\begin{aligned} \text{area} &= (P_{\text{AB}} - P_{\text{CD}})\Delta V \\ &= \left[(1.50\times10^{6}\ \text{N/m}^{2}) - (2.00\times10^{5}\ \text{N/m}^{2})\right](5.00\times10^{-4}\ \text{m}^{3}) \\ &= 650\ \text{J}. \end{aligned}$$

Thus,

Equation:

$$\text{area} = 650\ \text{J} = W.$$

Discussion

The result, as anticipated, is that the area inside the closed loop equals the work done. The area is often easier to calculate than is the work done along each path. It is also convenient to visualize the area inside different curves on PV diagrams in order to see which processes might produce the most work. Recall that work can be done to the system, or by the system, depending on the sign of W. A positive W is work that is done by the system on the outside environment; a negative W represents work done by the environment on the system.

[link](a) shows two other important processes on a PV diagram. For comparison, both are shown starting from the same point A. The upper curve ending at point B is an **isothermal** process—that is, one in which temperature is kept constant. If the gas behaves like an ideal gas, as is often the case, and if no phase change occurs, then $\mathrm{PV}=\mathrm{nRT}$. Since T is constant, PV is a constant for an isothermal process. We ordinarily expect the temperature of a gas to decrease as it expands, and so we correctly suspect that heat transfer must occur from the surroundings to the gas to keep the temperature constant during an isothermal expansion. To show this more rigorously for the special case of a monatomic ideal gas, we note that the average kinetic energy of an atom in such a gas is given by

Equation:

$$\frac{1}{2}mv^2=\frac{3}{2}\mathrm{kT}.$$

The kinetic energy of the atoms in a monatomic ideal gas is its only form of internal energy, and so its total internal energy U is

Equation:

$$U=N\frac{1}{2}mv^2=\frac{3}{2}\mathrm{NkT},\ \text{(monatomic ideal gas)},$$

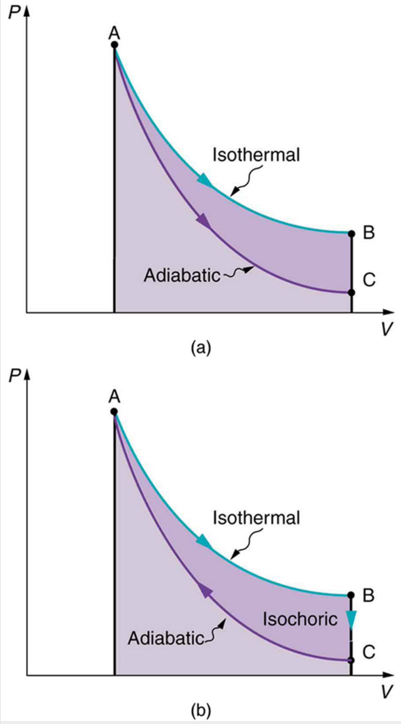
where N is the number of atoms in the gas. This relationship means that the internal energy of an ideal monatomic gas is constant during an isothermal process—that is, $\Delta U=0$. If the internal energy does not change, then the net heat transfer into the gas must equal the net work done by the gas. That is, because $\Delta U=Q-W=0$ here, $Q=W$. We must have just enough heat transfer to replace the work done. An isothermal process is inherently slow, because heat transfer occurs continuously to keep the gas temperature constant at all times and must be allowed to spread through the gas so that there are no hot or cold regions.

Also shown in [link](a) is a curve AC for an **adiabatic** process, defined to be one in which there is no heat transfer—that is, $Q = 0$. Processes that are nearly adiabatic can be achieved either by using very effective insulation or by performing the process so fast that there is little time for heat transfer. Temperature must decrease during an adiabatic expansion process, since work is done at the expense of internal energy:

Equation:

$$U = \frac{3}{2}\text{NkT}.$$

(You might have noted that a gas released into atmospheric pressure from a pressurized cylinder is substantially colder than the gas in the cylinder.) In fact, because $Q = 0$, $\Delta U =- W$ for an adiabatic process. Lower temperature results in lower pressure along the way, so that curve AC is lower than curve AB, and less work is done. If the path ABCA could be followed by cooling the gas from B to C at constant volume (isochorically), [link](b), there would be a net work output.



(a) The upper curve is an isothermal process ($\Delta T = 0$), whereas the lower curve is an adiabatic process ($Q = 0$). Both start from the same point A, but the isothermal process does more work than the adiabatic because heat transfer into the gas takes place to keep its temperature constant. This keeps the pressure higher all along the isothermal path than along the adiabatic path, producing more work. The adiabatic path thus ends up with a lower pressure and temperature at point C, even though the final volume is the same as for the isothermal process. (b) The cycle ABCA produces a net work output.

Reversible Processes

Both isothermal and adiabatic processes such as shown in [link] are reversible in principle. A **reversible process** is one in which both the system and its environment can return to exactly the states they were in by following the reverse path. The reverse isothermal and adiabatic paths are BA and CA, respectively. Real macroscopic processes are never exactly reversible. In the previous examples, our system is a gas (like that in [link]), and its environment is the piston, cylinder, and the rest of the universe. If there are any energy-dissipating mechanisms, such as friction or turbulence, then heat transfer to the environment occurs for either direction of the piston. So, for example, if the path BA is followed and there is friction, then the gas will be returned to its original state but the environment will not—it will have been heated in both directions. Reversibility requires the direction of heat transfer to reverse for the reverse path. Since dissipative mechanisms cannot be completely eliminated, real processes cannot be reversible.

There must be reasons that real macroscopic processes cannot be reversible. We can imagine them going in reverse. For example, heat transfer occurs spontaneously from hot to cold and never spontaneously the reverse. Yet it would not violate the first law of thermodynamics for this to happen. In fact, all spontaneous processes, such as bubbles bursting, never go in reverse. There is a second thermodynamic law that forbids them from going in reverse. When we study this law, we will learn something about nature and also find that such a law limits the efficiency of heat engines. We will find that heat engines with the greatest possible theoretical efficiency would have to use reversible processes, and even they cannot convert all heat transfer into doing work. [link] summarizes the simpler thermodynamic processes and their definitions.

Isobaric	Constant pressure $$W = P\Delta V$$
Isochoric	Constant volume $$W = 0$$
Isothermal	Constant temperature $$Q = W$$
Adiabatic	No heat transfer $$Q = 0$$

Summary of Simple Thermodynamic Processes

Note:
PhET Explorations: States of Matter
Watch different types of molecules form a solid, liquid, or gas. Add or remove heat and watch the phase change. Change the temperature or volume of a container and see a pressure-temperature diagram respond in real time. Relate the interaction potential to the forces between molecules.
https://phet.colorado.edu/sims/html/states-of-matter/latest/states-of-matter_en.html

Section Summary

- One of the important implications of the first law of thermodynamics is that machines can be harnessed to do work that humans previously did by hand or by external energy supplies such as running water or the heat of the Sun. A machine that uses heat transfer to do work is known as a heat engine.

- There are several simple processes, used by heat engines, that flow from the first law of thermodynamics. Among them are the isobaric, isochoric, isothermal and adiabatic processes.
- These processes differ from one another based on how they affect pressure, volume, temperature, and heat transfer.
- If the work done is performed on the outside environment, work (W) will be a positive value. If the work done is done to the heat engine system, work (W) will be a negative value.
- Some thermodynamic processes, including isothermal and adiabatic processes, are reversible in theory; that is, both the thermodynamic system and the environment can be returned to their initial states. However, because of loss of energy owing to the second law of thermodynamics, complete reversibility does not work in practice.

Conceptual Questions

Exercise:

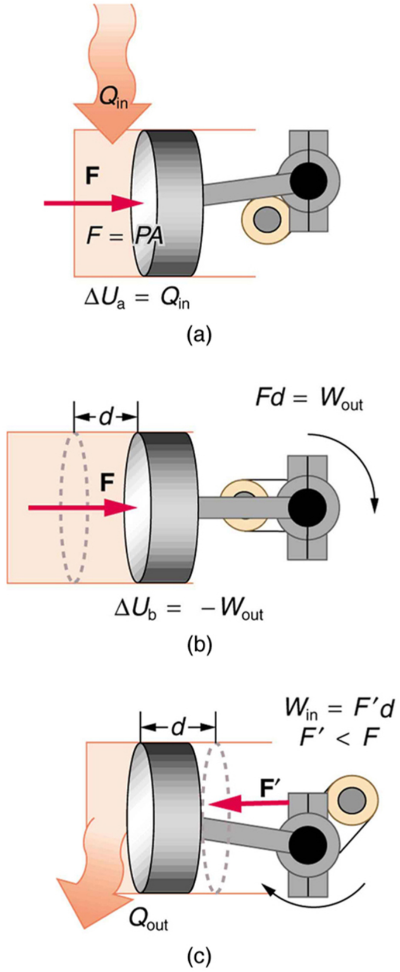
Problem:

A great deal of effort, time, and money has been spent in the quest for the so-called perpetual-motion machine, which is defined as a hypothetical machine that operates or produces useful work indefinitely and/or a hypothetical machine that produces more work or energy than it consumes. Explain, in terms of heat engines and the first law of thermodynamics, why or why not such a machine is likely to be constructed.

Exercise:

Problem:

One method of converting heat transfer into doing work is for heat transfer into a gas to take place, which expands, doing work on a piston, as shown in the figure below. (a) Is the heat transfer converted directly to work in an isobaric process, or does it go through another form first? Explain your answer. (b) What about in an isothermal process? (c) What about in an adiabatic process (where heat transfer occurred prior to the adiabatic process)?



Exercise:

Problem:

Would the previous question make any sense for an isochoric process? Explain your answer.

Exercise:

Problem:

We ordinarily say that $\Delta U = 0$ for an isothermal process. Does this assume no phase change takes place? Explain your answer.

Exercise:

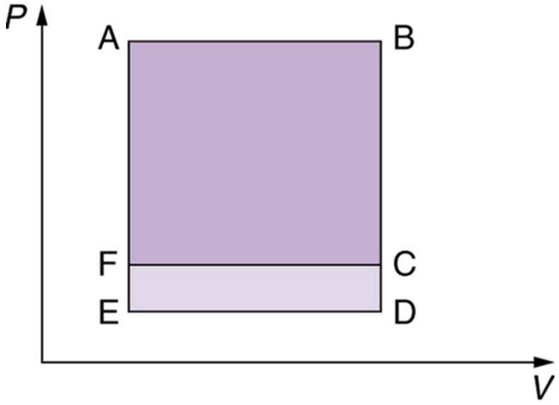
Problem:

The temperature of a rapidly expanding gas decreases. Explain why in terms of the first law of thermodynamics. (Hint: Consider whether the gas does work and whether heat transfer occurs rapidly into the gas through conduction.)

Exercise:

Problem:

Which cyclical process represented by the two closed loops, ABCFA and ABDEA, on the PV diagram in the figure below produces the greatest *net* work? Is that process also the one with the smallest work input required to return it to point A? Explain your responses.



The two cyclical processes shown on this PV diagram start with and return the system to the conditions at point A, but they follow

different paths and produce
different amounts of work.

Exercise:

Problem:

A real process may be nearly adiabatic if it occurs over a very short time. How does the short time span help the process to be adiabatic?

Exercise:

Problem:

It is unlikely that a process can be isothermal unless it is a very slow process. Explain why. Is the same true for isobaric and isochoric processes? Explain your answer.

Problem Exercises

Exercise:

Problem:

A car tire contains $0.0380\ \text{m}^3$ of air at a pressure of $2.20\times10^5\ \text{N/m}^2$ (about 32 psi). How much more internal energy does this gas have than the same volume has at zero gauge pressure (which is equivalent to normal atmospheric pressure)?

Solution:

$6.77 \times 10^3\ \text{J}$

Exercise:

Problem:

A helium-filled toy balloon has a gauge pressure of 0.200 atm and a volume of 10.0 L. How much greater is the internal energy of the helium in the balloon than it would be at zero gauge pressure?

Exercise:

Problem:

Steam to drive an old-fashioned steam locomotive is supplied at a constant gauge pressure of $1.75 \times 10^6 \text{ N/m}^2$ (about 250 psi) to a piston with a 0.200-m radius. (a) By calculating $P\Delta V$, find the work done by the steam when the piston moves 0.800 m. Note that this is the net work output, since gauge pressure is used. (b) Now find the amount of work by calculating the force exerted times the distance traveled. Is the answer the same as in part (a)?

Solution:

(a) $W = P\Delta V = 1.76 \times 10^5 \text{ J}$

(b) $W = \text{Fd} = 1.76 \times 10^5 \text{ J}$. Yes, the answer is the same.

Exercise:

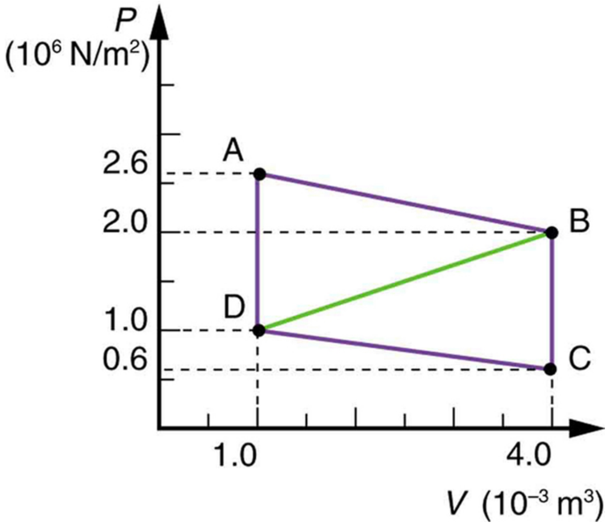
Problem:

A hand-driven tire pump has a piston with a 2.50-cm diameter and a maximum stroke of 30.0 cm. (a) How much work do you do in one stroke if the average gauge pressure is $2.40 \times 10^5 \text{ N/m}^2$ (about 35 psi)? (b) What average force do you exert on the piston, neglecting friction and gravitational force?

Exercise:

Problem:

Calculate the net work output of a heat engine following path ABCDA in the figure below.



Solution:

$W = 4.5 \times 10^3 \text{ J}$

Exercise:

Problem:

What is the net work output of a heat engine that follows path ABDA in the figure above, with a straight line from B to D? Why is the work output less than for path ABCDA? Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.

Exercise:

Problem: Unreasonable Results

What is wrong with the claim that a cyclical heat engine does 4.00 kJ of work on an input of 24.0 kJ of heat transfer while 16.0 kJ of heat transfers to the environment?

Solution:

W is not equal to the difference between the heat input and the heat output.

Exercise:

Problem:

(a) A cyclical heat engine, operating between temperatures of 450° C and 150° C produces 4.00 MJ of work on a heat transfer of 5.00 MJ into the engine. How much heat transfer occurs to the environment? (b) What is unreasonable about the engine? (c) Which premise is unreasonable?

Exercise:

Problem: Construct Your Own Problem

Consider a car's gasoline engine. Construct a problem in which you calculate the maximum efficiency this engine can have. Among the things to consider are the effective hot and cold reservoir temperatures. Compare your calculated efficiency with the actual efficiency of car engines.

Exercise:

Problem: Construct Your Own Problem

Consider a car trip into the mountains. Construct a problem in which you calculate the overall efficiency of the car for the trip as a ratio of kinetic and potential energy gained to fuel consumed. Compare this efficiency to the thermodynamic efficiency quoted for gasoline engines and discuss why the thermodynamic efficiency is so much greater. Among the factors to be considered are the gain in altitude and speed, the mass of the car, the distance traveled, and typical fuel economy.

Glossary

heat engine
: a machine that uses heat transfer to do work

isobaric process
: constant-pressure process in which a gas does work

isochoric process

a constant-volume process

isothermal process

a constant-temperature process

adiabatic process

a process in which no heat transfer takes place

reversible process

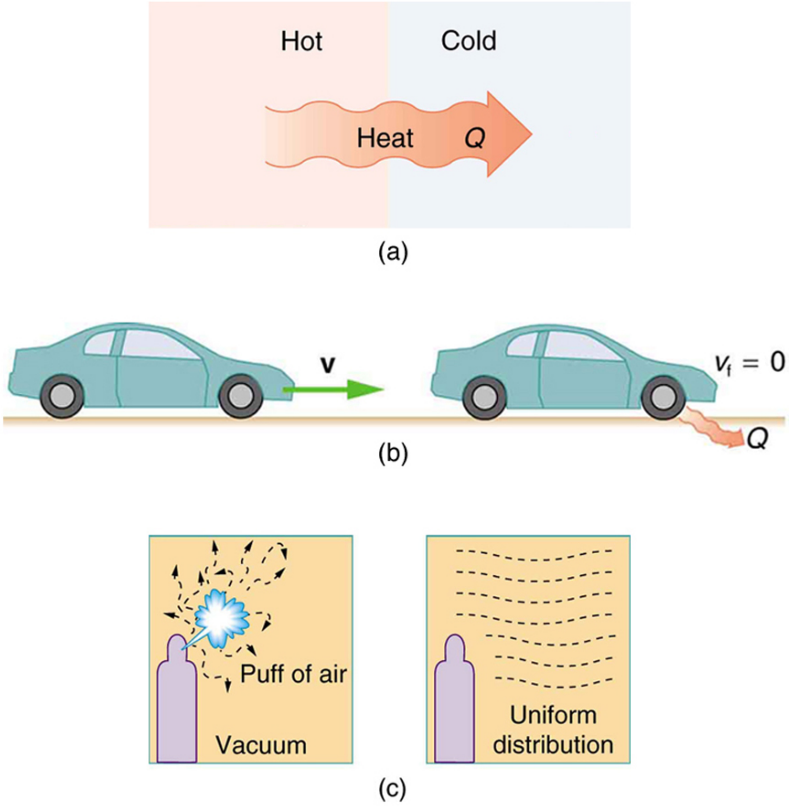
a process in which both the heat engine system and the external environment theoretically can be returned to their original states

Introduction to the Second Law of Thermodynamics: Heat Engines and Their Efficiency

- State the expressions of the second law of thermodynamics.
- Calculate the efficiency and carbon dioxide emission of a coal-fired electricity plant, using second law characteristics.
- Describe and define the Otto cycle.

These ice floes melt during the Arctic summer. Some of them refreeze in the winter, but the second law of thermodynamics predicts that it would be extremely unlikely for the water molecules contained in these particular floes to reform the distinctive alligator-like shape they formed when the picture was taken in the summer of 2009. (credit: Patrick Kelley, U.S. Coast Guard, U.S. Geological Survey)

The second law of thermodynamics deals with the direction taken by spontaneous processes. Many processes occur spontaneously in one direction only—that is, they are irreversible, under a given set of conditions. Although irreversibility is seen in day-to-day life—a broken glass does not resume its original state, for instance—complete irreversibility is a statistical statement that cannot be seen during the lifetime of the universe. More precisely, an **irreversible process** is one that depends on path. If the process can go in only one direction, then the reverse path differs fundamentally and the process cannot be reversible. For example, as noted in the previous section, heat involves the transfer of energy from higher to lower temperature. A cold object in contact with a hot one never gets colder, transferring heat to the hot object and making it hotter. Furthermore, mechanical energy, such as kinetic energy, can be completely converted to thermal energy by friction, but the reverse is impossible. A hot stationary object never spontaneously cools off and starts moving. Yet another example is the expansion of a puff of gas introduced into one corner of a vacuum chamber. The gas expands to fill the chamber, but it never regroups in the corner. The random motion of the gas molecules could take them all back to the corner, but this is never observed to happen. (See [link].)



Examples of one-way processes in nature.

(a) Heat transfer occurs spontaneously from hot to cold and not from cold to hot. (b) The brakes of this car convert its kinetic energy to heat transfer to the environment. The reverse process is impossible. (c) The burst of gas let into this vacuum chamber quickly expands to uniformly fill every part of the chamber. The random motions of the gas molecules will never return them to the corner.

The fact that certain processes never occur suggests that there is a law forbidding them to occur. The first law of thermodynamics would allow them to occur—none of those processes violate conservation of energy. The law that forbids these processes is called the second law of thermodynamics. We shall see that the second law can be stated in many ways that may seem different, but which in fact are equivalent. Like all natural laws, the second law of thermodynamics gives insights into nature, and its several statements imply that it is broadly applicable, fundamentally affecting many apparently disparate processes.

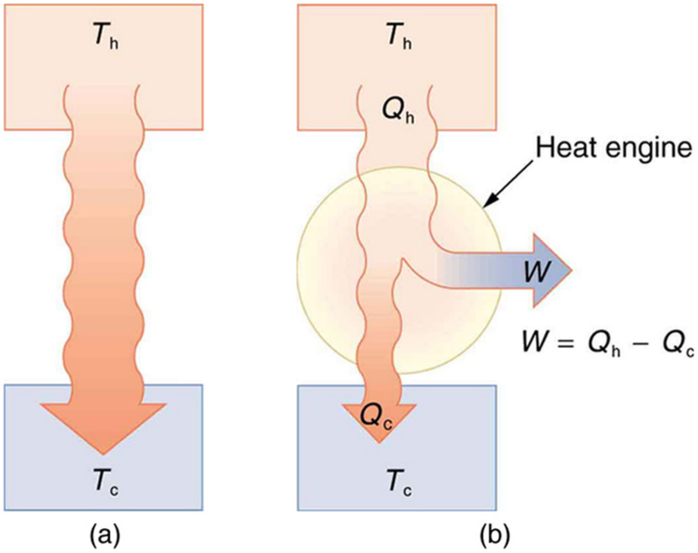
The already familiar direction of heat transfer from hot to cold is the basis of our first version of the **second law of thermodynamics**.

Note:
The Second Law of Thermodynamics (first expression)
Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction.

Another way of stating this: It is impossible for any process to have as its sole result heat transfer from a cooler to a hotter object.

Heat Engines

Now let us consider a device that uses heat transfer to do work. As noted in the previous section, such a device is called a heat engine, and one is shown schematically in [link](b). Gasoline and diesel engines, jet engines, and steam turbines are all heat engines that do work by using part of the heat transfer from some source. Heat transfer from the hot object (or hot reservoir) is denoted as Q_{h}, while heat transfer into the cold object (or cold reservoir) is Q_{c}, and the work done by the engine is W. The temperatures of the hot and cold reservoirs are T_{h} and T_{c}, respectively.



(a) Heat transfer occurs spontaneously from a hot object to a cold one, consistent with the second law of thermodynamics. (b) A heat engine, represented here by a circle, uses part of the heat transfer to do work. The hot and cold objects are called the hot and cold reservoirs. Q_{h} is the heat transfer out of the hot reservoir, W is the work output, and Q_{c} is the heat transfer into the cold reservoir.

Because the hot reservoir is heated externally, which is energy intensive, it is important that the work is done as efficiently as possible. In fact, we would like W to equal Q_{h}, and for there to be no heat transfer to the environment ($Q_{\text{c}} = 0$). Unfortunately, this is impossible. The **second law of thermodynamics** also states, with regard to using heat transfer to do work (the second expression of the second law):

Note:
The Second Law of Thermodynamics (second expression)
It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.

A **cyclical process** brings a system, such as the gas in a cylinder, back to its original state at the end of every cycle. Most heat engines, such as reciprocating piston engines and rotating turbines, use cyclical processes. The second law, just stated in its second form, clearly states that such engines cannot have perfect conversion of heat transfer into work done. Before going into the underlying reasons for the limits on converting heat transfer into work, we need to explore the relationships among W, Q_{h}, and Q_{c}, and to define the efficiency of a cyclical heat engine. As noted, a cyclical process brings the system back to its original condition at the end of every cycle. Such a system’s internal energy U is the same at the beginning and end of every cycle—that is, $\Delta U = 0$. The first law of thermodynamics states that

Equation:

$$\Delta U = Q - W,$$

where Q is the *net* heat transfer during the cycle ($Q = Q_{\text{h}} - Q_{\text{c}}$) and W is the net work done by the system. Since $\Delta U = 0$ for a complete cycle, we

have
Equation:

$$0 = Q - W,$$

so that
Equation:

$$W = Q.$$

Thus the net work done by the system equals the net heat transfer into the system, or
Equation:

$$W = Q_{\text{h}} - Q_{\text{c}}\ \text{(cyclical process)},$$

just as shown schematically in [link](b). The problem is that in all processes, there is some heat transfer Q_{c} to the environment—and usually a very significant amount at that.

In the conversion of energy to work, we are always faced with the problem of getting less out than we put in. We define *conversion efficiency* Eff to be the ratio of useful work output to the energy input (or, in other words, the ratio of what we get to what we spend). In that spirit, we define the efficiency of a heat engine to be its net work output W divided by heat transfer to the engine Q_{h}; that is,
Equation:

$$\text{Eff} = \frac{W}{Q_{\text{h}}}.$$

Since $W = Q_{\text{h}} - Q_{\text{c}}$ in a cyclical process, we can also express this as
Equation:

$$\text{Eff} = \frac{Q_\text{h} - Q_\text{c}}{Q_\text{h}} = 1 - \frac{Q_\text{c}}{Q_\text{h}} \text{ (cyclical process)},$$

making it clear that an efficiency of 1, or 100%, is possible only if there is no heat transfer to the environment ($Q_\text{c} = 0$). Note that all Qs are positive. The direction of heat transfer is indicated by a plus or minus sign. For example, Q_c is out of the system and so is preceded by a minus sign.

Example:

Daily Work Done by a Coal-Fired Power Station, Its Efficiency and Carbon Dioxide Emissions

A coal-fired power station is a huge heat engine. It uses heat transfer from burning coal to do work to turn turbines, which are used to generate electricity. In a single day, a large coal power station has 2.50×10^{14} J of heat transfer from coal and 1.48×10^{14} J of heat transfer into the environment. (a) What is the work done by the power station? (b) What is the efficiency of the power station? (c) In the combustion process, the following chemical reaction occurs: $\text{C} + \text{O}_2 \rightarrow \text{CO}_2$. This implies that every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of carbon dioxide into the atmosphere. Assuming that 1 kg of coal can provide 2.5×10^6 J of heat transfer upon combustion, how much CO_2 is emitted per day by this power plant?

Strategy for (a)

We can use $W = Q_\text{h} - Q_\text{c}$ to find the work output W, assuming a cyclical process is used in the power station. In this process, water is boiled under pressure to form high-temperature steam, which is used to run steam turbine-generators, and then condensed back to water to start the cycle again.

Solution for (a)

Work output is given by:

Equation:

$$W = Q_\text{h} - Q_\text{c}.$$

Substituting the given values:

Equation:

$$\begin{aligned} W &= 2.50{\times}10^{14}\ \text{J}–1.48{\times}10^{14}\ \text{J} \\ &= 1.02 \times 10^{14}\ \text{J}. \end{aligned}$$

Strategy for (b)

The efficiency can be calculated with $\text{Eff} = \frac{W}{Q_\text{h}}$ since Q_h is given and work W was found in the first part of this example.

Solution for (b)

Efficiency is given by: $\text{Eff} = \frac{W}{Q_\text{h}}$. The work W was just found to be 1.02×10^{14} J, and Q_h is given, so the efficiency is

Equation:

$$\begin{aligned} Eff &= \frac{1.02{\times}10^{14}\ \text{J}}{2.50{\times}10^{14}\ \text{J}} \\ &= 0.408\text{, or }40.8\% \end{aligned}$$

Strategy for (c)

The daily consumption of coal is calculated using the information that each day there is $2.50{\times}10^{14}$ J of heat transfer from coal. In the combustion process, we have $\text{C} + \text{O}_2 \rightarrow \text{CO}_2$. So every 12 kg of coal puts 12 kg + 16 kg + 16 kg = 44 kg of CO_2 into the atmosphere.

Solution for (c)

The daily coal consumption is

Equation:

$$\frac{2.50{\times}10^{14}\ \text{J}}{2.50{\times}10^{6}\ \text{J/kg}} = 1.0{\times}10^{8}\ \text{kg}.$$

Assuming that the coal is pure and that all the coal goes toward producing carbon dioxide, the carbon dioxide produced per day is

Equation:

$$1.0{\times}10^{8}\ \text{kg coal}{\times}\frac{44\ \text{kg CO}_2}{12\ \text{kg coal}} = 3.7{\times}10^{8}\ \text{kg CO}_2.$$

This is 370,000 metric tons of CO_2 produced every day.

Discussion

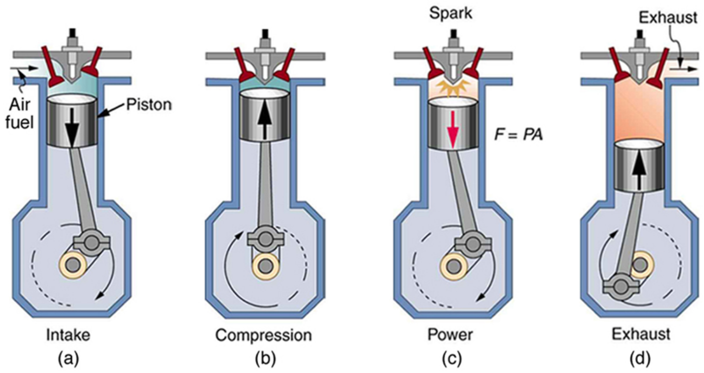
If all the work output is converted to electricity in a period of one day, the average power output is 1180 MW (this is left to you as an end-of-chapter problem). This value is about the size of a large-scale conventional power plant. The efficiency found is acceptably close to the value of 42% given for coal power stations. It means that fully 59.2% of the energy is heat transfer to the environment, which usually results in warming lakes, rivers, or the ocean near the power station, and is implicated in a warming planet generally. While the laws of thermodynamics limit the efficiency of such plants—including plants fired by nuclear fuel, oil, and natural gas—the heat transfer to the environment could be, and sometimes is, used for heating homes or for industrial processes. The generally low cost of energy has not made it economical to make better use of the waste heat transfer from most heat engines. Coal-fired power plants produce the greatest amount of CO_2 per unit energy output (compared to natural gas or oil), making coal the least efficient fossil fuel.

With the information given in [link], we can find characteristics such as the efficiency of a heat engine without any knowledge of how the heat engine operates, but looking further into the mechanism of the engine will give us greater insight. [link] illustrates the operation of the common four-stroke gasoline engine. The four steps shown complete this heat engine’s cycle, bringing the gasoline-air mixture back to its original condition.

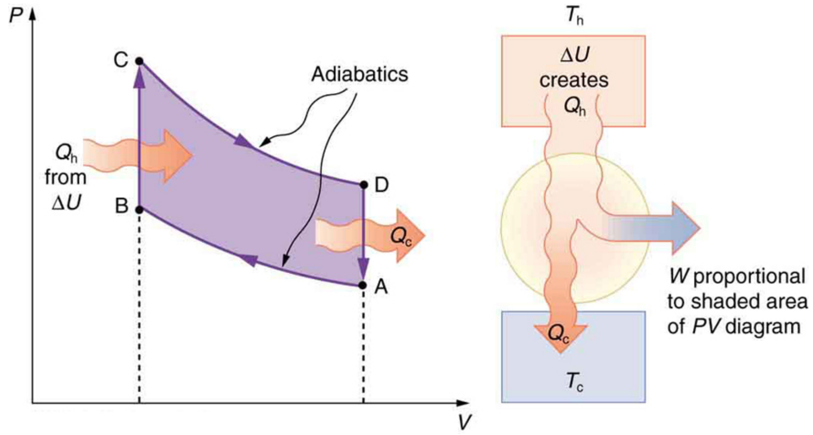
The **Otto cycle** shown in [link](a) is used in four-stroke internal combustion engines, although in fact the true Otto cycle paths do not correspond exactly to the strokes of the engine.

The adiabatic process AB corresponds to the nearly adiabatic compression stroke of the gasoline engine. In both cases, work is done on the system (the gas mixture in the cylinder), increasing its temperature and pressure. Along path BC of the Otto cycle, heat transfer Q_h into the gas occurs at constant volume, causing a further increase in pressure and temperature. This process corresponds to burning fuel in an internal combustion engine, and takes place so rapidly that the volume is nearly constant. Path CD in the Otto cycle is an adiabatic expansion that does work on the outside world, just as the power stroke of an internal combustion engine does in its nearly adiabatic expansion. The work done by the system along path CD is greater than the work done on the system along path AB, because the pressure is greater, and so there is a net work output. Along path DA in the Otto cycle, heat transfer Q_c from the gas at constant volume reduces its temperature and pressure, returning it to its original state. In an internal combustion engine, this process corresponds to the exhaust of hot gases and the intake of an air-gasoline mixture at a considerably lower temperature. In both cases, heat transfer into the environment occurs along this final path.

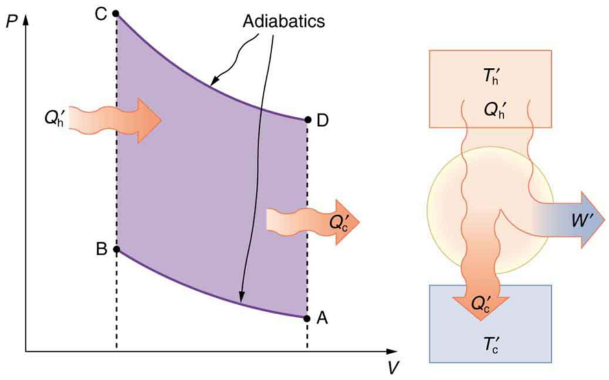
The net work done by a cyclical process is the area inside the closed path on a PV diagram, such as that inside path ABCDA in [link]. Note that in every imaginable cyclical process, it is absolutely necessary for heat transfer from the system to occur in order to get a net work output. In the Otto cycle, heat transfer occurs along path DA. If no heat transfer occurs, then the return path is the same, and the net work output is zero. The lower the temperature on the path AB, the less work has to be done to compress the gas. The area inside the closed path is then greater, and so the engine does more work and is thus more efficient. Similarly, the higher the temperature along path CD, the more work output there is. (See [link].) So efficiency is related to the temperatures of the hot and cold reservoirs. In the next section, we shall see what the absolute limit to the efficiency of a heat engine is, and how it is related to temperature.



In the four-stroke internal combustion gasoline engine, heat transfer into work takes place in the cyclical process shown here. The piston is connected to a rotating crankshaft, which both takes work out of and does work on the gas in the cylinder. (a) Air is mixed with fuel during the intake stroke. (b) During the compression stroke, the air-fuel mixture is rapidly compressed in a nearly adiabatic process, as the piston rises with the valves closed. Work is done on the gas. (c) The power stroke has two distinct parts. First, the air-fuel mixture is ignited, converting chemical potential energy into thermal energy almost instantaneously, which leads to a great increase in pressure. Then the piston descends, and the gas does work by exerting a force through a distance in a nearly adiabatic process. (d) The exhaust stroke expels the hot gas to prepare the engine for another cycle, starting again with the intake stroke.



PV diagram for a simplified Otto cycle, analogous to that employed in an internal combustion engine. Point A corresponds to the start of the compression stroke of an internal combustion engine. Paths AB and CD are adiabatic and correspond to the compression and power strokes of an internal combustion engine, respectively. Paths BC and DA are isochoric and accomplish similar results to the ignition and exhaust-intake portions, respectively, of the internal combustion engine’s cycle. Work is done on the gas along path AB, but more work is done by the gas along path CD, so that there is a net work output.



This Otto cycle produces a greater work output than the one in [link], because the starting temperature of path CD is higher and the starting temperature of path AB is lower. The area inside the loop is greater, corresponding to greater net work output.

Section Summary

- The two expressions of the second law of thermodynamics are: (i) Heat transfer occurs spontaneously from higher- to lower-temperature bodies but never spontaneously in the reverse direction; and (ii) It is impossible in any system for heat transfer from a reservoir to completely convert to work in a cyclical process in which the system returns to its initial state.
- Irreversible processes depend on path and do not return to their original state. Cyclical processes are processes that return to their original state at the end of every cycle.
- In a cyclical process, such as a heat engine, the net work done by the system equals the net heat transfer into the system, or $W = Q_{\text{h}} - Q_{\text{c}}$, where Q_{h} is the heat transfer from the hot object (hot reservoir), and Q_{c} is the heat transfer into the cold object (cold reservoir).

- Efficiency can be expressed as $\text{Eff} = \frac{W}{Q_{\text{h}}}$, the ratio of work output divided by the amount of energy input.
- The four-stroke gasoline engine is often explained in terms of the Otto cycle, which is a repeating sequence of processes that convert heat into work.

Conceptual Questions

Exercise:

Problem:

Imagine you are driving a car up Pike’s Peak in Colorado. To raise a car weighing 1000 kilograms a distance of 100 meters would require about a million joules. You could raise a car 12.5 kilometers with the energy in a gallon of gas. Driving up Pike's Peak (a mere 3000-meter climb) should consume a little less than a quart of gas. But other considerations have to be taken into account. Explain, in terms of efficiency, what factors may keep you from realizing your ideal energy use on this trip.

Exercise:

Problem:

Is a temperature difference necessary to operate a heat engine? State why or why not.

Exercise:

Problem:

Definitions of efficiency vary depending on how energy is being converted. Compare the definitions of efficiency for the human body and heat engines. How does the definition of efficiency in each relate to the type of energy being converted into doing work?

Exercise:

Problem:

Why—other than the fact that the second law of thermodynamics says reversible engines are the most efficient—should heat engines employing reversible processes be more efficient than those employing irreversible processes? Consider that dissipative mechanisms are one cause of irreversibility.

Problem Exercises

Exercise:

Problem:

A certain heat engine does 10.0 kJ of work and 8.50 kJ of heat transfer occurs to the environment in a cyclical process. (a) What was the heat transfer into this engine? (b) What was the engine’s efficiency?

Solution:

(a) $18.5\ \text{kJ}$

(b) 54.1%

Exercise:

Problem:

With $2.56 \times 10^{6}\ \text{J}$ of heat transfer into this engine, a given cyclical heat engine can do only $1.50 \times 10^{5}\ \text{J}$ of work. (a) What is the engine’s efficiency? (b) How much heat transfer to the environment takes place?

Exercise:

Problem:

(a) What is the work output of a cyclical heat engine having a 22.0% efficiency and $6.00\times10^{9}\text{ J}$ of heat transfer into the engine? (b) How much heat transfer occurs to the environment?

Solution:

(a) $1.32\times10^{9}\text{ J}$

(b) $4.68\times10^{9}\text{ J}$

Exercise:

Problem:

(a) What is the efficiency of a cyclical heat engine in which 75.0 kJ of heat transfer occurs to the environment for every 95.0 kJ of heat transfer into the engine? (b) How much work does it produce for 100 kJ of heat transfer into the engine?

Exercise:

Problem:

The engine of a large ship does $2.00\times10^{8}\text{ J}$ of work with an efficiency of 5.00%. (a) How much heat transfer occurs to the environment? (b) How many barrels of fuel are consumed, if each barrel produces $6.00\times10^{9}\text{ J}$ of heat transfer when burned?

Solution:

(a) $3.80\times10^{9}\text{ J}$

(b) 0.667 barrels

Exercise:

Problem:

(a) How much heat transfer occurs to the environment by an electrical power station that uses $1.25\times10^{14}\text{ J}$ of heat transfer into the engine with an efficiency of 42.0%? (b) What is the ratio of heat transfer to the environment to work output? (c) How much work is done?

Exercise:

Problem:

Assume that the turbines at a coal-powered power plant were upgraded, resulting in an improvement in efficiency of 3.32%. Assume that prior to the upgrade the power station had an efficiency of 36% and that the heat transfer into the engine in one day is still the same at $2.50\times10^{14}\text{ J}$. (a) How much more electrical energy is produced due to the upgrade? (b) How much less heat transfer occurs to the environment due to the upgrade?

Solution:

(a) $8.30\times10^{12}\text{ J}$, which is 3.32% of $2.50\times10^{14}\text{ J}$.

(b) –$8.30\times10^{12}\text{ J}$, where the negative sign indicates a reduction in heat transfer to the environment.

Exercise:

Problem:

This problem compares the energy output and heat transfer to the environment by two different types of nuclear power stations—one with the normal efficiency of 34.0%, and another with an improved efficiency of 40.0%. Suppose both have the same heat transfer into the engine in one day, $2.50 \times 10^{14}\ \text{J}$. (a) How much more electrical energy is produced by the more efficient power station? (b) How much less heat transfer occurs to the environment by the more efficient power station? (One type of more efficient nuclear power station, the gas-cooled reactor, has not been reliable enough to be economically feasible in spite of its greater efficiency.)

Glossary

irreversible process
any process that depends on path direction

second law of thermodynamics
heat transfer flows from a hotter to a cooler object, never the reverse, and some heat energy in any process is lost to available work in a cyclical process

cyclical process
a process in which the path returns to its original state at the end of every cycle

Otto cycle
a thermodynamic cycle, consisting of a pair of adiabatic processes and a pair of isochoric processes, that converts heat into work, e.g., the four-stroke engine cycle of intake, compression, ignition, and exhaust

Carnot’s Perfect Heat Engine: The Second Law of Thermodynamics Restated

- Identify a Carnot cycle.
- Calculate maximum theoretical efficiency of a nuclear reactor.
- Explain how dissipative processes affect the ideal Carnot engine.

This novelty toy, known as the drinking bird, is an example of Carnot’s engine. It contains methylene chloride (mixed with a dye) in the abdomen, which boils at a very low temperature—about 100°F. To operate, one gets the bird’s head wet. As the water evaporates, fluid moves up into the head, causing the bird to become top-heavy and dip forward back into the water. This cools down the methylene chloride in the head, and it moves back into the abdomen, causing the bird to become bottom heavy and tip up. Except for a very small input of energy—the original head-wetting—the bird becomes a perpetual motion machine of sorts. (credit: Arabesk.nl, Wikimedia Commons)

We know from the second law of thermodynamics that a heat engine cannot be 100% efficient, since there must always be some heat transfer Q_c to the environment, which is often called waste heat. How efficient, then, can a heat engine be? This question was answered at a theoretical level in 1824 by a young French engineer, Sadi Carnot (1796–1832), in his study of the then-emerging heat engine technology crucial to the Industrial Revolution. He devised a theoretical cycle, now called the **Carnot cycle**, which is the most efficient cyclical process possible. The second law of thermodynamics can be restated in terms of the Carnot cycle, and so what Carnot actually discovered was this fundamental law. Any heat engine employing the Carnot cycle is called a **Carnot engine**.

What is crucial to the Carnot cycle—and, in fact, defines it—is that only reversible processes are used. Irreversible processes involve dissipative factors, such as friction and turbulence. This increases heat transfer Q_c to the environment and reduces the efficiency of the engine. Obviously, then, reversible processes are superior.

Note:
Carnot Engine
Stated in terms of reversible processes, the **second law of thermodynamics** has a third form:
A Carnot engine operating between two given temperatures has the greatest possible efficiency of any heat engine operating between these two temperatures. Furthermore, all engines employing only reversible processes have this same maximum efficiency when operating between the same given temperatures.

[link] shows the PV diagram for a Carnot cycle. The cycle comprises two isothermal and two adiabatic processes. Recall that both isothermal and adiabatic processes are, in principle, reversible.

Carnot also determined the efficiency of a perfect heat engine—that is, a Carnot engine. It is always true that the efficiency of a cyclical heat engine is given by:

Equation:

$$\text{Eff} = \frac{Q_\text{h} - Q_\text{c}}{Q_\text{h}} = 1 - \frac{Q_\text{c}}{Q_\text{h}}.$$

What Carnot found was that for a perfect heat engine, the ratio Q_c/Q_h equals the ratio of the absolute temperatures of the heat reservoirs. That is, $Q_\text{c}/Q_\text{h} = T_\text{c}/T_\text{h}$ for a Carnot engine, so that the maximum or **Carnot efficiency** Eff_C is given by

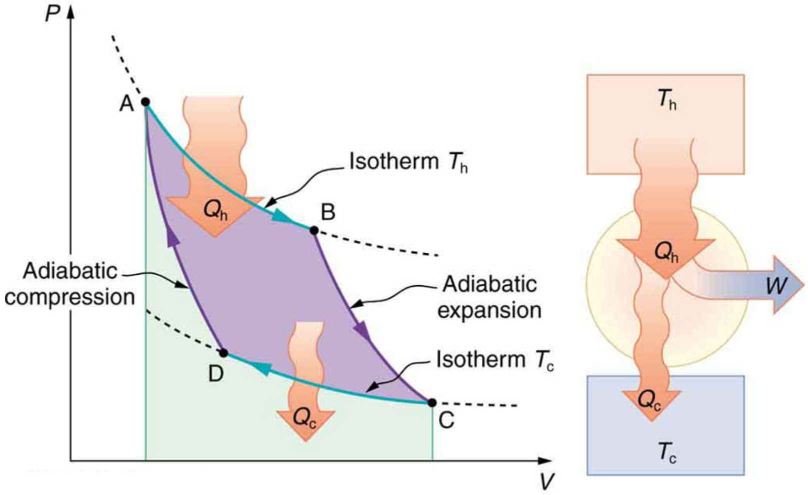
Equation:

$$Eff_\text{C} = 1 - \frac{T_\text{c}}{T_\text{h}},$$

where T_h and T_c are in kelvins (or any other absolute temperature scale). No real heat engine can do as well as the Carnot efficiency—an actual efficiency of about 0.7 of this maximum is usually the best that can be accomplished. But the ideal Carnot engine, like the drinking bird above, while a fascinating novelty, has zero power. This makes it unrealistic for any applications.

Carnot’s interesting result implies that 100% efficiency would be possible only if $T_\text{c} = 0\text{ K}$ —that is, only if the cold reservoir were at absolute zero, a practical and theoretical impossibility. But the physical implication is this —the only way to have all heat transfer go into doing work is to remove *all* thermal energy, and this requires a cold reservoir at absolute zero.

It is also apparent that the greatest efficiencies are obtained when the ratio T_c/T_h is as small as possible. Just as discussed for the Otto cycle in the previous section, this means that efficiency is greatest for the highest possible temperature of the hot reservoir and lowest possible temperature of the cold reservoir. (This setup increases the area inside the closed loop on the PV diagram; also, it seems reasonable that the greater the temperature difference, the easier it is to divert the heat transfer to work.) The actual reservoir temperatures of a heat engine are usually related to the type of heat source and the temperature of the environment into which heat transfer occurs. Consider the following example.

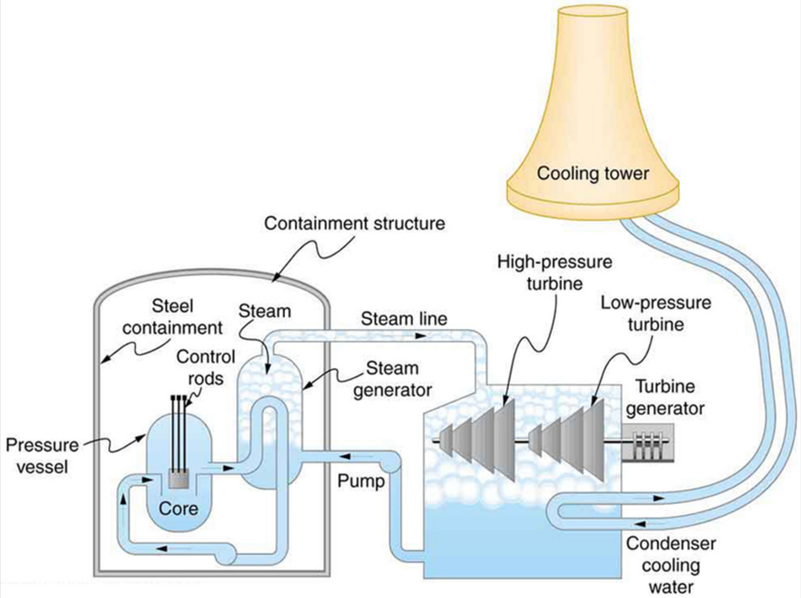


PV diagram for a Carnot cycle, employing only reversible isothermal and adiabatic processes. Heat transfer Q_{h} occurs into the working substance during the isothermal path AB, which takes place at constant temperature T_{h}. Heat transfer Q_{c} occurs out of the working substance during the isothermal path CD, which takes place at constant temperature T_{c}. The net work output W equals the area inside the path ABCDA. Also shown is a schematic of a Carnot engine operating between hot and cold reservoirs at temperatures T_{h} and T_{c}. Any heat engine using reversible processes and operating between these two temperatures will have the same maximum efficiency as the Carnot engine.

Example:

Maximum Theoretical Efficiency for a Nuclear Reactor

A nuclear power reactor has pressurized water at $300^{o}C$. (Higher temperatures are theoretically possible but practically not, due to limitations with materials used in the reactor.) Heat transfer from this water is a complex process (see [link]). Steam, produced in the steam generator, is used to drive the turbine-generators. Eventually the steam is condensed to water at $27^{o}C$ and then heated again to start the cycle over. Calculate the maximum theoretical efficiency for a heat engine operating between these two temperatures.



Schematic diagram of a pressurized water nuclear reactor and the steam turbines that convert work into electrical energy. Heat exchange is used to generate steam, in part to avoid contamination of the generators with radioactivity. Two turbines are used because this is less expensive than operating a single generator that produces the same amount of electrical energy. The steam is condensed to liquid before being returned to the heat exchanger, to keep exit steam pressure low and aid the flow of steam through the turbines (equivalent to using a lower-temperature cold reservoir). The considerable energy associated with condensation must be dissipated into the local environment; in this example, a cooling tower is used so there is no direct heat transfer to an aquatic environment. (Note that the water going to the cooling tower does not come into contact with the steam flowing over the turbines.)

Strategy

Since temperatures are given for the hot and cold reservoirs of this heat engine, $Eff_{\text{C}} = 1 - \frac{T_{\text{c}}}{T_{\text{h}}}$ can be used to calculate the Carnot (maximum theoretical) efficiency. Those temperatures must first be converted to kelvins.

Solution

The hot and cold reservoir temperatures are given as 300^{o}C and 27.0^{o}C, respectively. In kelvins, then, $T_{\text{h}} = 573\text{ K}$ and $T_{\text{c}} = 300\text{ K}$, so that the maximum efficiency is

Equation:

$$Eff_{\text{C}} = 1 - \frac{T_{\text{c}}}{T_{\text{h}}}.$$

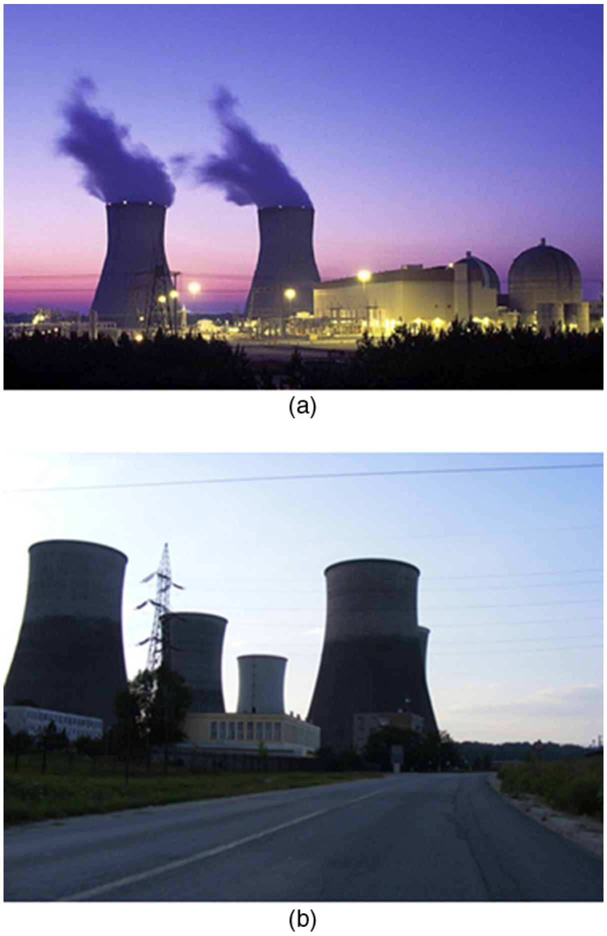
Thus,

Equation:

$$\begin{array}{lll} Eff_{\text{C}} & = & 1 - \frac{300\text{ K}}{573\text{ K}} \\ & = & 0.476\text{, or }47.6\%. \end{array}$$

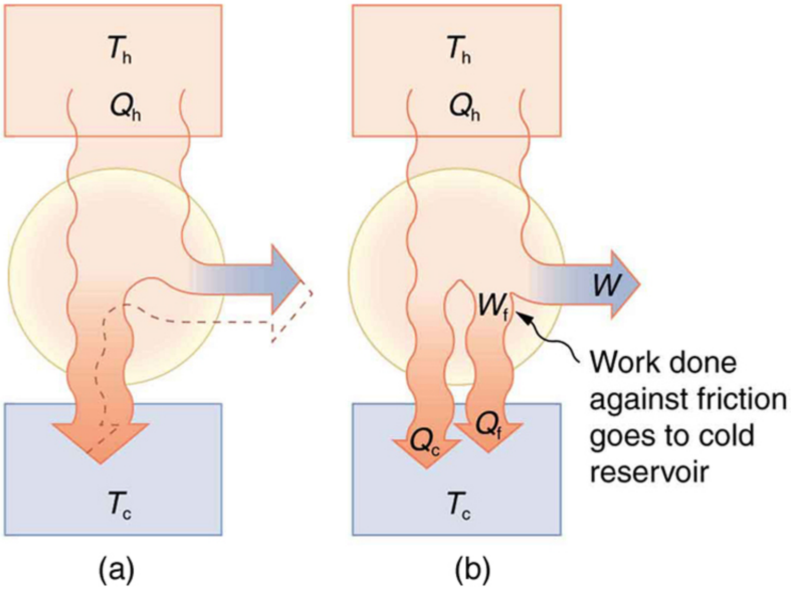
Discussion

A typical nuclear power station’s actual efficiency is about 35%, a little better than 0.7 times the maximum possible value, a tribute to superior engineering. Electrical power stations fired by coal, oil, and natural gas have greater actual efficiencies (about 42%), because their boilers can reach higher temperatures and pressures. The cold reservoir temperature in any of these power stations is limited by the local environment. [link] shows (a) the exterior of a nuclear power station and (b) the exterior of a coal-fired power station. Both have cooling towers into which water from the condenser enters the tower near the top and is sprayed downward, cooled by evaporation.



(a) A nuclear power station (credit: BlatantWorld.com) and (b) a coal-fired power station. Both have cooling towers in which water evaporates into the environment, representing Q_c. The nuclear reactor, which supplies Q_h, is housed inside the dome-shaped containment buildings. (credit: Robert & Mihaela Vicol, publicphoto.org)

Since all real processes are irreversible, the actual efficiency of a heat engine can never be as great as that of a Carnot engine, as illustrated in [link](a). Even with the best heat engine possible, there are always dissipative processes in peripheral equipment, such as electrical transformers or car transmissions. These further reduce the overall efficiency by converting some of the engine's work output back into heat transfer, as shown in [link](b).



Real heat engines are less efficient than Carnot engines. (a) Real engines use irreversible processes, reducing the heat transfer to work. Solid lines represent the actual process; the dashed lines are what a Carnot engine would do between the same two reservoirs. (b) Friction and other dissipative processes in the output mechanisms of a heat engine convert some of its work output into heat transfer to the environment.

Section Summary

- The Carnot cycle is a theoretical cycle that is the most efficient cyclical process possible. Any engine using the Carnot cycle, which uses only reversible processes (adiabatic and isothermal), is known as a Carnot engine.
- Any engine that uses the Carnot cycle enjoys the maximum theoretical efficiency.
- While Carnot engines are ideal engines, in reality, no engine achieves Carnot's theoretical maximum efficiency, since dissipative processes, such as friction, play a role. Carnot cycles without heat loss may be possible at absolute zero, but this has never been seen in nature.

Conceptual Questions

Exercise:

Problem:

Think about the drinking bird at the beginning of this section ([link]). Although the bird enjoys the theoretical maximum efficiency possible, if left to its own devices over time, the bird will cease "drinking." What are some of the dissipative processes that might cause the bird's motion to cease?

Exercise:

Problem:

Can improved engineering and materials be employed in heat engines to reduce heat transfer into the environment? Can they eliminate heat transfer into the environment entirely?

Exercise:

Problem:

Does the second law of thermodynamics alter the conservation of energy principle?

Problem Exercises

Exercise:

Problem:

A certain gasoline engine has an efficiency of 30.0%. What would the hot reservoir temperature be for a Carnot engine having that efficiency, if it operates with a cold reservoir temperature of 200°C?

Solution:

403°C

Exercise:

Problem:

A gas-cooled nuclear reactor operates between hot and cold reservoir temperatures of 700°C and 27.0°C. (a) What is the maximum efficiency of a heat engine operating between these temperatures? (b) Find the ratio of this efficiency to the Carnot efficiency of a standard nuclear reactor (found in [link]).

Exercise:

Problem:

(a) What is the hot reservoir temperature of a Carnot engine that has an efficiency of 42.0% and a cold reservoir temperature of 27.0°C? (b) What must the hot reservoir temperature be for a real heat engine that achieves 0.700 of the maximum efficiency, but still has an efficiency of 42.0% (and a cold reservoir at 27.0°C)? (c) Does your answer imply practical limits to the efficiency of car gasoline engines?

Solution:

(a) 244ºC

(b) 477ºC

(c)Yes, since automobiles engines cannot get too hot without overheating, their efficiency is limited.

Exercise:

Problem:

Steam locomotives have an efficiency of 17.0% and operate with a hot steam temperature of 425ºC. (a) What would the cold reservoir temperature be if this were a Carnot engine? (b) What would the maximum efficiency of this steam engine be if its cold reservoir temperature were 150ºC?

Exercise:

Problem:

Practical steam engines utilize 450ºC steam, which is later exhausted at 270ºC. (a) What is the maximum efficiency that such a heat engine can have? (b) Since 270ºC steam is still quite hot, a second steam engine is sometimes operated using the exhaust of the first. What is the maximum efficiency of the second engine if its exhaust has a temperature of 150ºC? (c) What is the overall efficiency of the two engines? (d) Show that this is the same efficiency as a single Carnot engine operating between 450ºC and 150ºC. Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.

Solution:

(a) $Eff_1 = 1 - \frac{T_{\text{c,1}}}{T_{\text{h,1}}} = 1 - \frac{543\text{ K}}{723\text{ K}} = 0.249 \text{ or } 24.9\%$

(b) $Eff_2 = 1 - \frac{423\text{ K}}{543\text{ K}} = 0.221 \text{ or } 22.1\%$

(c) $Eff_1 = 1 - \frac{T_{\text{c},1}}{T_{\text{h},1}} \Rightarrow T_{\text{c},1} = T_{\text{h},1}(1, -, eff_1)$
similarly, $T_{\text{c},2} = T_{\text{h},2}(1 - Eff_2)$
using $T_{\text{h},2} = T_{\text{c},1}$ in above equation gives
$T_{\text{c},2} = T_{\text{h},1}(1 - Eff_1)(1 - Eff_2) \equiv T_{\text{h},1}(1 - Eff_{\text{overall}})$
$(1 - Eff_{\text{overall}}) = (1 - Eff_1)(1 - Eff_2)$
$Eff_{\text{overall}} = 1 - (1 - 0.249)(1 - 0.221) = 41.5\%$

(d) $Eff_{\text{overall}} = 1 - \frac{423\text{ K}}{723\text{ K}} = 0.415 \text{ or } 41.5\%$

Exercise:

Problem:

A coal-fired electrical power station has an efficiency of 38%. The temperature of the steam leaving the boiler is 550^{o}C. What percentage of the maximum efficiency does this station obtain? (Assume the temperature of the environment is 20^{o}C.)

Exercise:

Problem:

Would you be willing to financially back an inventor who is marketing a device that she claims has 25 kJ of heat transfer at 600 K, has heat transfer to the environment at 300 K, and does 12 kJ of work? Explain your answer.

Solution:

The heat transfer to the cold reservoir is $Q_{\text{c}} = Q_{\text{h}} - W = 25\text{ kJ} - 12\text{ kJ} = 13\text{ kJ}$, so the efficiency is $Eff = 1 - \frac{Q_{\text{c}}}{Q_{\text{h}}} = 1 - \frac{13\text{ kJ}}{25\text{ kJ}} = 0.48$. The Carnot efficiency is $Eff_{\text{C}} = 1 - \frac{T_{\text{c}}}{T_{\text{h}}} = 1 - \frac{300\text{ K}}{600\text{ K}} = 0.50$. The actual efficiency is 96% of the Carnot efficiency, which is much higher than the best-ever achieved of about 70%, so her scheme is likely to be fraudulent.

Exercise:

Problem: **Unreasonable Results**

(a) Suppose you want to design a steam engine that has heat transfer to the environment at $270^{\circ}C$ and has a Carnot efficiency of 0.800. What temperature of hot steam must you use? (b) What is unreasonable about the temperature? (c) Which premise is unreasonable?

Exercise:

Problem: **Unreasonable Results**

Calculate the cold reservoir temperature of a steam engine that uses hot steam at $450^{\circ}C$ and has a Carnot efficiency of 0.700. (b) What is unreasonable about the temperature? (c) Which premise is unreasonable?

Solution:

(a) $-56.3^{\circ}C$

(b) The temperature is too cold for the output of a steam engine (the local environment). It is below the freezing point of water.

(c) The assumed efficiency is too high.

Glossary

Carnot cycle
a cyclical process that uses only reversible processes, the adiabatic and isothermal processes

Carnot engine
a heat engine that uses a Carnot cycle

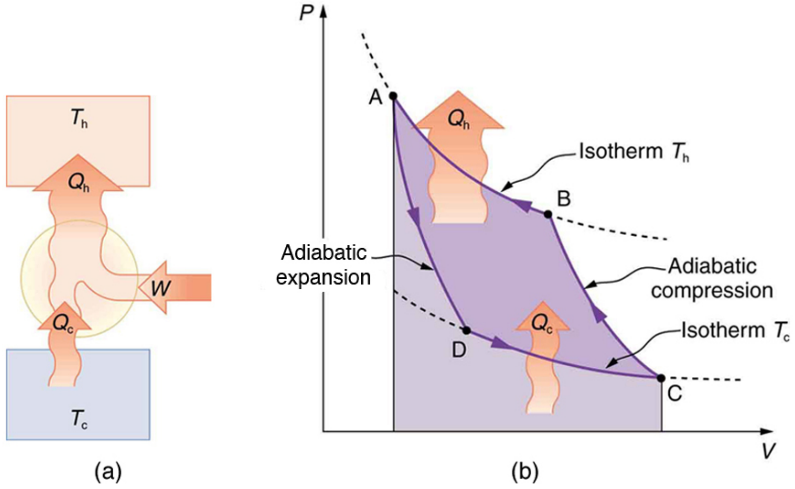
Carnot efficiency
the maximum theoretical efficiency for a heat engine

Applications of Thermodynamics: Heat Pumps and Refrigerators

- Describe the use of heat engines in heat pumps and refrigerators.
- Demonstrate how a heat pump works to warm an interior space.
- Explain the differences between heat pumps and refrigerators.
- Calculate a heat pump's coefficient of performance.

Almost every home contains a refrigerator. Most people don't realize they are also sharing their homes with a heat pump. (credit: Id1337x, Wikimedia Commons)

Heat pumps, air conditioners, and refrigerators utilize heat transfer from cold to hot. They are heat engines run backward. We say backward, rather than reverse, because except for Carnot engines, all heat engines, though they can be run backward, cannot truly be reversed. Heat transfer occurs from a cold reservoir Q_{c} and into a hot one. This requires work input W, which is also converted to heat transfer. Thus the heat transfer to the hot reservoir is $Q_{\mathrm{h}} = Q_{\mathrm{c}} + W$. (Note that Q_{h}, Q_{c}, and W are positive, with their directions indicated on schematics rather than by sign.) A heat pump's mission is for heat transfer Q_{h} to occur into a warm environment, such as a home in the winter. The mission of air conditioners and refrigerators is for heat transfer Q_{c} to occur from a cool environment, such as chilling a room or keeping food at lower temperatures than the environment. (Actually, a heat pump can be used both to heat and cool a space. It is essentially an air conditioner and a heating unit all in one. In this section we will concentrate on its heating mode.)

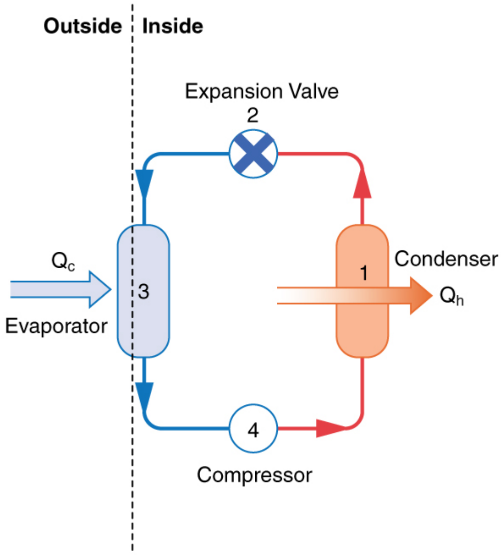


Heat pumps, air conditioners, and refrigerators are heat engines operated backward. The one shown here is based on a Carnot (reversible) engine. (a) Schematic diagram showing heat transfer from a cold reservoir to a warm reservoir with a heat pump. The directions of W, Q_{h}, and Q_{c} are opposite what they would be in a heat engine. (b) PV diagram for a Carnot cycle similar to that in [link] but reversed, following path ADCBA. The area inside the loop is negative, meaning there is a net work input. There is heat transfer Q_{c} into the system from a cold reservoir along path DC, and heat transfer Q_{h} out of the system into a hot reservoir along path BA.

Heat Pumps

The great advantage of using a heat pump to keep your home warm, rather than just burning fuel, is that a heat pump supplies $Q_{\text{h}} = Q_{\text{c}} + W$. Heat transfer is from the outside air, even at a temperature below freezing, to the indoor space. You only pay for W, and you get an additional heat transfer of Q_{c} from the outside at no cost; in many cases, at least twice as much energy is transferred to the heated space as is used to run the heat pump. When you burn fuel to keep warm, you pay for all of it. The disadvantage is that the work input (required by the second law of thermodynamics) is sometimes more expensive than simply burning fuel, especially if the work is done by electrical energy.

The basic components of a heat pump in its heating mode are shown in [link]. A working fluid such as a non-CFC refrigerant is used. In the outdoor coils (the evaporator), heat transfer Q_{c} occurs to the working fluid from the cold outdoor air, turning it into a gas.



A simple heat pump has four basic components: (1) condenser, (2) expansion valve, (3) evaporator, and (4) compressor. In the heating mode, heat transfer Q_c occurs to the working fluid in the evaporator (3) from the colder outdoor air, turning it into a gas. The electrically driven compressor (4) increases the temperature and pressure of the gas and forces it into the condenser coils (1) inside the heated space. Because the temperature of the gas is higher than the temperature in the room, heat transfer from the gas to the room occurs as the gas condenses to a liquid. The working fluid is then cooled as it flows back through an expansion valve (2) to the outdoor evaporator coils.

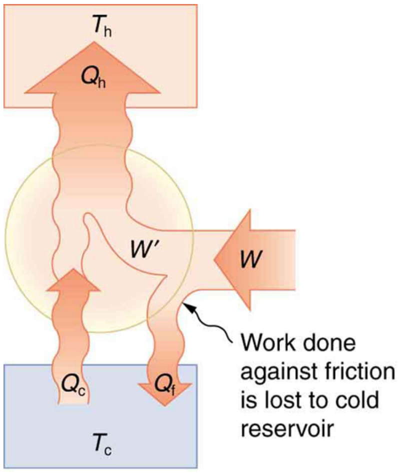
The electrically driven compressor (work input W) raises the temperature and pressure of the gas and forces it into the condenser coils that are inside the heated space. Because the temperature of the gas is higher than the temperature inside the room, heat transfer to the room occurs and the gas condenses to a liquid. The liquid then flows back through a pressure-reducing valve to the outdoor evaporator coils, being cooled through expansion. (In a cooling cycle, the evaporator and condenser coils exchange roles and the flow direction of the fluid is reversed.)

The quality of a heat pump is judged by how much heat transfer Q_{h} occurs into the warm space compared with how much work input W is required. In the spirit of taking the ratio of what you get to what you spend, we define a **heat pump's coefficient of performance** (COP_{hp}) to be
Equation:

$$COP_{\mathrm{hp}} = \frac{Q_{\mathrm{h}}}{W}.$$

Since the efficiency of a heat engine is $Eff = W/Q_{\mathrm{h}}$, we see that $COP_{\mathrm{hp}} = 1/Eff$, an important and interesting fact. First, since the efficiency of any heat engine is less than 1, it means that COP_{hp} is always greater than 1—that is, a heat pump always has more heat transfer Q_{h} than work put into it. Second, it means that heat pumps work best when temperature differences are small. The efficiency of a perfect, or Carnot, engine is $Eff_{\mathrm{C}} = 1 - (T_{\mathrm{c}}/T_{\mathrm{h}})$; thus, the smaller the temperature difference, the smaller the efficiency and the greater the COP_{hp} (because $COP_{\mathrm{hp}} = 1/Eff$). In other words, heat pumps do not work as well in very cold climates as they do in more moderate climates.

Friction and other irreversible processes reduce heat engine efficiency, but they do *not* benefit the operation of a heat pump—instead, they reduce the work input by converting part of it to heat transfer back into the cold reservoir before it gets into the heat pump.



When a real heat engine is run backward, some of the intended work input (W) goes into heat transfer before it gets into the heat engine, thereby reducing its coefficient of performance COP_{hp}. In this figure, W’ represents the portion of W that goes into the heat pump, while the remainder of W is lost in the form of frictional heat (Q_f) to the cold reservoir. If all of W had gone into the heat pump, then Q_{h} would have been greater. The best heat pump uses adiabatic and isothermal processes, since, in theory, there would be no dissipative processes to reduce the heat transfer to the hot reservoir.

Example:

The Best *COP* hp of a Heat Pump for Home Use

A heat pump used to warm a home must employ a cycle that produces a working fluid at temperatures greater than typical indoor temperature so that heat transfer to the inside can take place. Similarly, it must produce a working fluid at temperatures that are colder than the outdoor temperature so that heat transfer occurs from outside. Its hot and cold reservoir temperatures therefore cannot be too close, placing a limit on its COP_{hp}. (See [link].) What is the best coefficient of performance possible for such a heat pump, if it has a hot reservoir temperature of 45.0^{o}C and a cold reservoir temperature of -15.0^{o}C?

Strategy

A Carnot engine reversed will give the best possible performance as a heat pump. As noted above, $COP_{\text{hp}}=1/Eff$, so that we need to first calculate the Carnot efficiency to solve this problem.

Solution

Carnot efficiency in terms of absolute temperature is given by**:**

Equation:

$$Eff_{\text{C}}=1-\frac{T_{\text{c}}}{T_{\text{h}}}.$$

The temperatures in kelvins are $T_{\text{h}} = 318\text{ K}$ and $T_{\text{c}} = 258\text{ K}$, so that
Equation:

$$Eff_{\text{C}} = 1 - \frac{258\text{ K}}{318\text{ K}} = 0.1887.$$

Thus, from the discussion above,
Equation:

$$COP_{\text{hp}} = \frac{1}{\text{Eff}} = \frac{1}{0.1887} = 5.30,$$

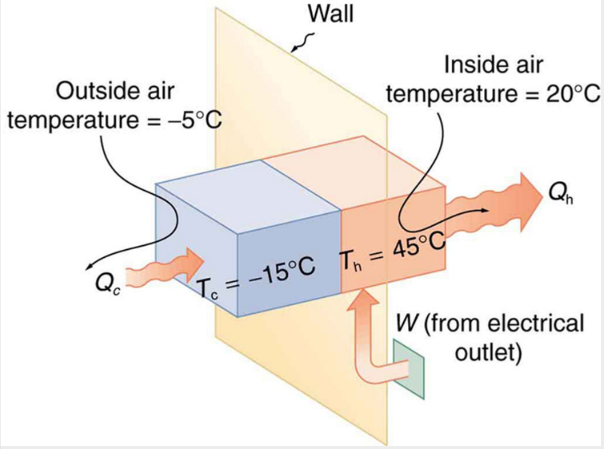
or
Equation:

$$COP_{\text{hp}} = \frac{Q_{\text{h}}}{W} = 5.30,$$

so that
Equation:

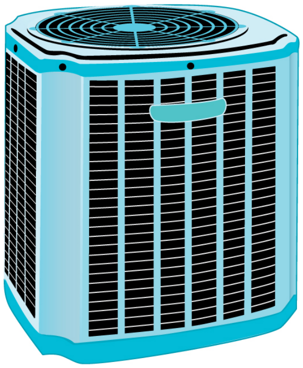
$$Q_{\text{h}} = 5.30\text{ W}.$$

Discussion
This result means that the heat transfer by the heat pump is 5.30 times as much as the work put into it. It would cost 5.30 times as much for the same heat transfer by an electric room heater as it does for that produced by this heat pump. This is not a violation of conservation of energy. Cold ambient air provides 4.3 J per 1 J of work from the electrical outlet.



Heat transfer from the outside to the inside, along with work done to run the pump, takes place in the heat pump of the example above**.** Note that the cold temperature produced by the heat pump is lower than the outside temperature, so that heat transfer into the working fluid occurs. The pump’s compressor produces a temperature greater than the indoor temperature in order for heat transfer into the house to occur.

Real heat pumps do not perform quite as well as the ideal one in the previous example; their values of COP_{hp} range from about 2 to 4. This range means that the heat transfer Q_{h} from the heat pumps is 2 to 4 times as great as the work W put into them. Their economical feasibility is still limited, however, since W is usually supplied by electrical energy that costs more per joule than heat transfer by burning fuels like natural gas. Furthermore, the initial cost of a heat pump is greater than that of many furnaces, so that a heat pump must last longer for its cost to be recovered. Heat pumps are most likely to be economically superior where winter temperatures are mild, electricity is relatively cheap, and other fuels are relatively expensive. Also, since they can cool as well as heat a space, they have advantages where cooling in summer months is also desired. Thus some of the best locations for heat pumps are in warm summer climates with cool winters. [link] shows a heat pump, called a “*reverse cycle*” or “*split-system cooler*” in some countries.



In hot weather, heat transfer occurs from air inside the room to air outside, cooling the room. In cool weather, heat transfer occurs from air outside to air inside, warming the room. This switching is achieved by reversing the direction of flow of the working fluid.

Air Conditioners and Refrigerators

Air conditioners and refrigerators are designed to cool something down in a warm environment. As with heat pumps, work input is required for heat transfer from cold to hot, and this is expensive. The quality of air conditioners and refrigerators is judged by how much heat transfer Q_{c} occurs from a cold environment compared with how much work input W is required. What is considered the benefit in a heat pump is considered waste heat in a refrigerator. We thus define the **coefficient of performance** (COP_{ref}) of an air conditioner or refrigerator to be

Equation:

$$COP_{\text{ref}} = \frac{Q_{\text{c}}}{W}.$$

Noting again that $Q_{\text{h}} = Q_{\text{c}} + W$, we can see that an air conditioner will have a lower coefficient of performance than a heat pump, because $COP_{\text{hp}} = Q_{\text{h}}/W$ and Q_{h} is greater than Q_{c}. In this module's Problems and Exercises, you will show that

Equation:

$$COP_{\text{ref}} = COP_{\text{hp}} - 1$$

for a heat engine used as either an air conditioner or a heat pump operating between the same two temperatures. Real air conditioners and refrigerators typically do remarkably well, having values of COP_{ref} ranging from 2 to 6.

These numbers are better than the COP_{hp} values for the heat pumps mentioned above, because the temperature differences are smaller, but they are less than those for Carnot engines operating between the same two temperatures.

A type of COP rating system called the "energy efficiency rating" (EER) has been developed. This rating is an example where non-SI units are still used and relevant to consumers. To make it easier for the consumer, Australia, Canada, New Zealand, and the U.S. use an Energy Star Rating out of 5 stars—the more stars, the more energy efficient the appliance. EERs are expressed in mixed units of British thermal units (Btu) per hour of heating or cooling divided by the power input in watts. Room air conditioners are readily available with EERs ranging from 6 to 12. Although not the same as the COPs just described, these EERs are good for comparison purposes—the greater the EER, the cheaper an air conditioner is to operate (but the higher its purchase price is likely to be).

The EER of an air conditioner or refrigerator can be expressed as
Equation:

$$EER = \frac{Q_{\text{c}}/t_1}{W/t_2},$$

where Q_{c} is the amount of heat transfer from a cold environment in British thermal units, t_1 is time in hours, W is the work input in joules, and t_2 is time in seconds.

Note:
Problem-Solving Strategies for Thermodynamics

1. *Examine the situation to determine whether heat, work, or internal energy are involved.* Look for any system where the primary methods of transferring energy are heat and work. Heat engines, heat pumps, refrigerators, and air conditioners are examples of such systems.

2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful. Maximum efficiency means a Carnot engine is involved. Efficiency is not the same as the coefficient of performance.
4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Be sure to distinguish heat transfer into a system from heat transfer out of the system, as well as work input from work output. In many situations, it is useful to determine the type of process, such as isothermal or adiabatic.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).*
6. *Substitute the known quantities along with their units into the appropriate equation and obtain numerical solutions complete with units.*
7. *Check the answer to see if it is reasonable: Does it make sense?* For example, efficiency is always less than 1, whereas coefficients of performance are greater than 1.

Section Summary

- An artifact of the second law of thermodynamics is the ability to heat an interior space using a heat pump. Heat pumps compress cold ambient air and, in so doing, heat it to room temperature without violation of conservation principles.
- To calculate the heat pump's coefficient of performance, use the equation $COP_{\text{hp}} = \frac{Q_{\text{h}}}{W}$.
- A refrigerator is a heat pump; it takes warm ambient air and expands it to chill it.

Conceptual Questions

Exercise:

Problem:

Explain why heat pumps do not work as well in very cold climates as they do in milder ones. Is the same true of refrigerators?

Exercise:

Problem:

In some Northern European nations, homes are being built without heating systems of any type. They are very well insulated and are kept warm by the body heat of the residents. However, when the residents are not at home, it is still warm in these houses. What is a possible explanation?

Exercise:

Problem:

Why do refrigerators, air conditioners, and heat pumps operate most cost-effectively for cycles with a small difference between T_h and T_c? (Note that the temperatures of the cycle employed are crucial to its COP.)

Exercise:

Problem:

Grocery store managers contend that there is *less* total energy consumption in the summer if the store is kept at a *low* temperature. Make arguments to support or refute this claim, taking into account that there are numerous refrigerators and freezers in the store.

Exercise:

Problem:

Can you cool a kitchen by leaving the refrigerator door open?

Problem Exercises

Exercise:

Problem:

What is the coefficient of performance of an ideal heat pump that has heat transfer from a cold temperature of -25.0°C to a hot temperature of 40.0°C?

Solution:

4.82

Exercise:

Problem:

Suppose you have an ideal refrigerator that cools an environment at -20.0°C and has heat transfer to another environment at 50.0°C. What is its coefficient of performance?

Exercise:

Problem:

What is the best coefficient of performance possible for a hypothetical refrigerator that could make liquid nitrogen at -200°C and has heat transfer to the environment at 35.0°C?

Solution:

0.311

Exercise:

Problem:

In a very mild winter climate, a heat pump has heat transfer from an environment at 5.00°C to one at 35.0°C. What is the best possible coefficient of performance for these temperatures? Explicitly show how you follow the steps in the Problem-Solving Strategies for Thermodynamics.

Exercise:

Problem:

(a) What is the best coefficient of performance for a heat pump that has a hot reservoir temperature of $50.0^{\mathrm{o}}\mathrm{C}$ and a cold reservoir temperature of $-20.0^{\mathrm{o}}\mathrm{C}$? (b) How much heat transfer occurs into the warm environment if $3.60 \times 10^{7}\ \mathrm{J}$ of work ($10.0\mathrm{kW} \cdot \mathrm{h}$) is put into it? (c) If the cost of this work input is $10.0\ \mathrm{cents/kW} \cdot \mathrm{h}$, how does its cost compare with the direct heat transfer achieved by burning natural gas at a cost of 85.0 cents per therm. (A therm is a common unit of energy for natural gas and equals $1.055 \times 10^{8}\ \mathrm{J}$.)

Solution:

(a) 4.61

(b) $1.66 \times 10^{8}\ \mathrm{J}$ or $3.97 \times 10^{4}\ \mathrm{kcal}$

(c) To transfer $1.66 \times 10^{8}\ \mathrm{J}$, heat pump costs $1.00, natural gas costs $1.34.

Exercise:

Problem:

(a) What is the best coefficient of performance for a refrigerator that cools an environment at $-30.0^{\mathrm{o}}\mathrm{C}$ and has heat transfer to another environment at $45.0^{\mathrm{o}}\mathrm{C}$? (b) How much work in joules must be done for a heat transfer of 4186 kJ from the cold environment? (c) What is the cost of doing this if the work costs 10.0 cents per $3.60 \times 10^{6}\ \mathrm{J}$ (a kilowatt-hour)? (d) How many kJ of heat transfer occurs into the warm environment? (e) Discuss what type of refrigerator might operate between these temperatures.

Exercise:

Problem:

Suppose you want to operate an ideal refrigerator with a cold temperature of -10.0°C, and you would like it to have a coefficient of performance of 7.00. What is the hot reservoir temperature for such a refrigerator?

Solution:

27.6°C

Exercise:

Problem:

An ideal heat pump is being considered for use in heating an environment with a temperature of 22.0°C. What is the cold reservoir temperature if the pump is to have a coefficient of performance of 12.0?

Exercise:

Problem:

A 4-ton air conditioner removes $5.06\times10^{7}\text{ J}$ (48,000 British thermal units) from a cold environment in 1.00 h. (a) What energy input in joules is necessary to do this if the air conditioner has an energy efficiency rating (EER) of 12.0? (b) What is the cost of doing this if the work costs 10.0 cents per $3.60\times10^{6}\text{ J}$ (one kilowatt-hour)? (c) Discuss whether this cost seems realistic. Note that the energy efficiency rating (EER) of an air conditioner or refrigerator is defined to be the number of British thermal units of heat transfer from a cold environment per hour divided by the watts of power input.

Solution:

(a) $1.44\times10^{7}\text{ J}$

(b) 40 cents

(c) This cost seems quite realistic; it says that running an air conditioner all day would cost $9.59 (if it ran continuously).

Exercise:

Problem:

Show that the coefficients of performance of refrigerators and heat pumps are related by $COP_{\text{ref}} = COP_{\text{hp}} - 1$.

Start with the definitions of the COP s and the conservation of energy relationship between Q_{h}, Q_{c}, and W.

Glossary

heat pump

a machine that generates heat transfer from cold to hot

coefficient of performance

for a heat pump, it is the ratio of heat transfer at the output (the hot reservoir) to the work supplied; for a refrigerator or air conditioner, it is the ratio of heat transfer from the cold reservoir to the work supplied

Entropy and the Second Law of Thermodynamics: Disorder and the Unavailability of Energy

- Define entropy and calculate the increase of entropy in a system with reversible and irreversible processes.
- Explain the expected fate of the universe in entropic terms.
- Calculate the increasing disorder of a system.

The ice in this drink is slowly melting. Eventually the liquid will reach thermal equilibrium, as predicted by the second law of thermodynamics. (credit: Jon Sullivan, PDPhoto.org)

There is yet another way of expressing the second law of thermodynamics. This version relates to a concept called **entropy**. By examining it, we shall see that the directions associated with the second law—heat transfer from hot to cold, for example—are related to the tendency in nature for systems to become disordered and for less energy to be available for use as work. The entropy of a system can in fact be shown to be a measure of its disorder and of the unavailability of energy to do work.

Note:
Making Connections: Entropy, Energy, and Work
Recall that the simple definition of energy is the ability to do work. Entropy is a measure of how much energy is not available to do work. Although all forms of energy are interconvertible, and all can be used to do work, it is not always possible, even in principle, to convert the entire available energy into work. That unavailable energy is of interest in thermodynamics, because the field of thermodynamics arose from efforts to convert heat to work.

We can see how entropy is defined by recalling our discussion of the Carnot engine. We noted that for a Carnot cycle, and hence for any reversible processes, $Q_{\mathrm{c}}/Q_{\mathrm{h}} = T_{\mathrm{c}}/T_{\mathrm{h}}$. Rearranging terms yields
Equation:

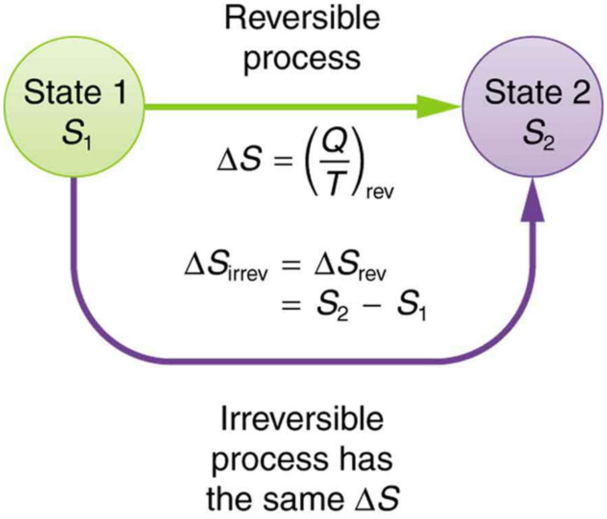
$$\frac{Q_{\mathrm{c}}}{T_{\mathrm{c}}} = \frac{Q_{\mathrm{h}}}{T_{\mathrm{h}}}$$

for any reversible process. Q_{c} and Q_{h} are absolute values of the heat transfer at temperatures T_{c} and T_{h}, respectively. This ratio of Q/T is defined to be the **change in entropy** ΔS for a reversible process,
Equation:

$$\Delta S = \left(\frac{Q}{T}\right)_{\mathrm{rev}},$$

where Q is the heat transfer, which is positive for heat transfer into and negative for heat transfer out of, and T is the absolute temperature at which the reversible process takes place. The SI unit for entropy is joules per kelvin (J/K). If temperature changes during the process, then it is usually a good approximation (for small changes in temperature) to take T to be the average temperature, avoiding the need to use integral calculus to find ΔS.

The definition of ΔS is strictly valid only for reversible processes, such as used in a Carnot engine. However, we can find ΔS precisely even for real, irreversible processes. The reason is that the entropy S of a system, like internal energy U, depends only on the state of the system and not how it reached that condition. Entropy is a property of state. Thus the change in entropy ΔS of a system between state 1 and state 2 is the same no matter how the change occurs. We just need to find or imagine a reversible process that takes us from state 1 to state 2 and calculate ΔS for that process. That will be the change in entropy for any process going from state 1 to state 2. (See [link].)


Reversible process
State 1 S_1
State 2 S_2
$\Delta S = \left(\frac{Q}{T}\right)_{\text{rev}}$
$\Delta S_{\text{irrev}} = \Delta S_{\text{rev}} = S_2 - S_1$
Irreversible process has the same ΔS

When a system goes from state 1 to state 2, its entropy changes by the same amount ΔS, whether a hypothetical reversible path is followed or a real irreversible path is taken.

Now let us take a look at the change in entropy of a Carnot engine and its heat reservoirs for one full cycle. The hot reservoir has a loss of entropy $\Delta S_{\text{h}} = -Q_{\text{h}}/T_{\text{h}}$, because heat transfer occurs out of it (remember that when heat transfers out, then Q has a negative sign). The cold reservoir has a gain of entropy $\Delta S_{\text{c}} = Q_{\text{c}}/T_{\text{c}}$, because heat transfer occurs into it. (We assume the reservoirs are sufficiently large that their temperatures are constant.) So the total change in entropy is

Equation:

$$\Delta S_{\text{tot}} = \Delta S_{\text{h}} + \Delta S_{\text{c}}.$$

Thus, since we know that $Q_{\text{h}}/T_{\text{h}} = Q_{\text{c}}/T_{\text{c}}$ for a Carnot engine,

Equation:

$$\Delta S_{\text{tot}}{=}{-}\frac{Q_{\text{h}}}{T_{\text{h}}} + \frac{Q_{\text{c}}}{T_{\text{c}}} = 0.$$

This result, which has general validity, means that *the total change in entropy for a system in any reversible process is zero.*

The entropy of various parts of the system may change, but the total change is zero. Furthermore, the system does not affect the entropy of its surroundings, since heat transfer between them does not occur. Thus the reversible process changes neither the total entropy of the system nor the entropy of its surroundings. Sometimes this is stated as follows: *Reversible processes do not affect the total entropy of the universe.* Real processes are not reversible, though, and they do change total entropy. We can, however, use hypothetical reversible processes to determine the value of entropy in real, irreversible processes. The following example illustrates this point.

Example:

Entropy Increases in an Irreversible (Real) Process

Spontaneous heat transfer from hot to cold is an irreversible process. Calculate the total change in entropy if 4000 J of heat transfer occurs from a hot reservoir at $T_{\text{h}} = 600\text{ K}(327^{\text{o}}\text{ C})$ to a cold reservoir at $T_{\text{c}} = 250\text{ K}(-23^{\text{o}}\text{ C})$, assuming there is no temperature change in either reservoir. (See [link].)

Strategy

How can we calculate the change in entropy for an irreversible process when $\Delta S_{\text{tot}} = \Delta S_{\text{h}} + \Delta S_{\text{c}}$ is valid only for reversible processes? Remember that the total change in entropy of the hot and cold reservoirs will be the same whether a reversible or irreversible process is involved in heat transfer from hot to cold. So we can calculate the change in entropy of the hot reservoir for a hypothetical reversible process in which 4000 J of heat transfer occurs from it; then we do the same for a hypothetical reversible process in which 4000 J of heat transfer occurs to the cold reservoir. This produces the same changes in the hot and cold reservoirs that would occur if the heat transfer were allowed to occur irreversibly between them, and so it also produces the same changes in entropy.

Solution

We now calculate the two changes in entropy using $\Delta S_{\text{tot}} = \Delta S_{\text{h}} + \Delta S_{\text{c}}$. First, for the heat transfer from the hot reservoir,

Equation:

$$\Delta S_{\text{h}} = \frac{-Q_{\text{h}}}{T_{\text{h}}} = \frac{-4000\text{ J}}{600\text{ K}} =- 6.67\text{ J/K}.$$

And for the cold reservoir,

Equation:

$$\Delta S_{\text{c}} = \frac{Q_{\text{c}}}{T_{\text{c}}} = \frac{4000\text{ J}}{250\text{ K}} = 16.0\text{ J/K}.$$

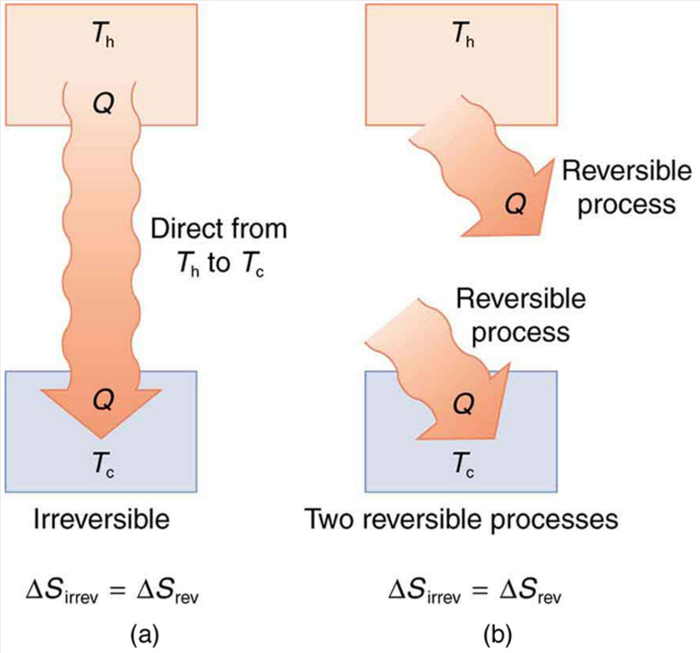
Thus the total is

Equation:

$$\begin{aligned} \Delta S_{\text{tot}} &= \Delta S_{\text{h}} + \Delta S_{\text{c}} \\ &= (- 6.67 + 16.0)\text{ J/K} \\ &= 9.33\text{ J/K}. \end{aligned}$$

Discussion

There is an *increase* in entropy for the system of two heat reservoirs undergoing this irreversible heat transfer. We will see that this means there is a loss of ability to do work with this transferred energy. Entropy has increased, and energy has become unavailable to do work.



(a) Heat transfer from a hot object to a cold one is an irreversible process that produces an overall increase in entropy. (b) The same final state and, thus, the same change in entropy is achieved for the objects if reversible heat transfer processes occur between the two objects whose temperatures are the same as the temperatures of the corresponding objects in the irreversible process.

It is reasonable that entropy increases for heat transfer from hot to cold. Since the change in entropy is Q/T, there is a larger change at lower temperatures. The decrease in entropy of the hot object is therefore less than the increase in entropy of the cold object, producing an overall increase, just as in the previous example. This result is very general:

There is an increase in entropy for any system undergoing an irreversible process.

With respect to entropy, there are only two possibilities: entropy is constant for a reversible process, and it increases for an irreversible process. There is a fourth version of **the second law of thermodynamics stated in terms of entropy**:

The total entropy of a system either increases or remains constant in any process; it never decreases.

For example, heat transfer cannot occur spontaneously from cold to hot, because entropy would decrease.

Entropy is very different from energy. Entropy is *not* conserved but increases in all real processes. Reversible processes (such as in Carnot engines) are the processes in which the most heat transfer to work takes place and are also the ones that keep entropy constant. Thus we are led to make a connection between entropy and the availability of energy to do work.

Entropy and the Unavailability of Energy to Do Work

What does a change in entropy mean, and why should we be interested in it? One reason is that entropy is directly related to the fact that not all heat transfer can be converted into work. The next example gives some indication of how an increase in entropy results in less heat transfer into work.

Example:
Less Work is Produced by a Given Heat Transfer When Entropy Change is Greater

(a) Calculate the work output of a Carnot engine operating between temperatures of 600 K and 100 K for 4000 J of heat transfer to the engine. (b) Now suppose that the 4000 J of heat transfer occurs first from the 600 K reservoir to a 250 K reservoir (without doing any work, and this produces the increase in entropy calculated above) before transferring into a Carnot engine operating between 250 K and 100 K. What work output is produced? (See [link].)

Strategy

In both parts, we must first calculate the Carnot efficiency and then the work output.

Solution (a)

The Carnot efficiency is given by

Equation:

$$Eff_{\text{C}} = 1 - \frac{T_{\text{c}}}{T_{\text{h}}}.$$

Substituting the given temperatures yields

Equation:

$$Eff_{\text{C}} = 1 - \frac{100\text{ K}}{600\text{ K}} = 0.833.$$

Now the work output can be calculated using the definition of efficiency for any heat engine as given by

Equation:

$$\text{Eff} = \frac{W}{Q_{\text{h}}}.$$

Solving for W and substituting known terms gives

Equation:

$$\begin{array}{rcl} W & = & Eff_{\text{C}}Q_{\text{h}} \\ & = & (0.833)(4000\text{ J}) = 3333\text{ J}. \end{array}$$

Solution (b)

Similarly,

Equation:

$$Eff\prime_{\text{C}} = 1 - \frac{T_{\text{c}}}{T\prime_{\text{c}}} = 1 - \frac{100\text{ K}}{250\text{ K}} = 0.600,$$

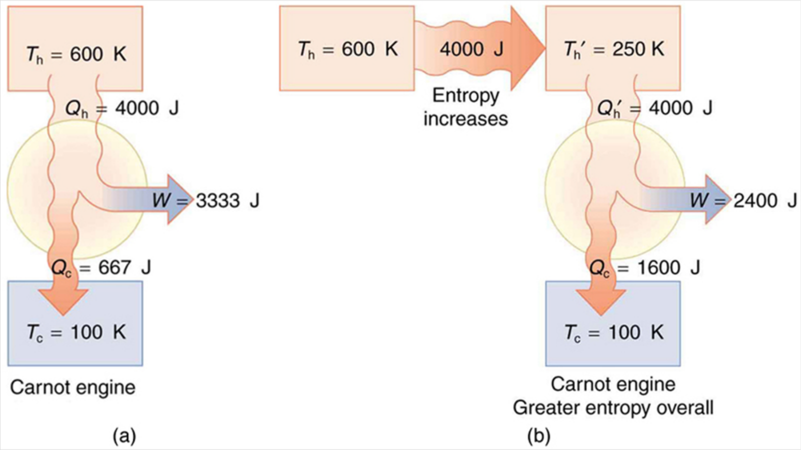
so that

Equation:

$$\begin{aligned} W &= Eff\prime_{\text{C}} Q_h \\ &= (0.600)(4000\text{ J}) = 2400\text{ J}. \end{aligned}$$

Discussion

There is 933 J less work from the same heat transfer in the second process. This result is important. The same heat transfer into two perfect engines produces different work outputs, because the entropy change differs in the two cases. In the second case, entropy is greater and less work is produced. Entropy is associated with the *un*availability of energy to do work.



(a) A Carnot engine working at between 600 K and 100 K has 4000 J of heat transfer and performs 3333 J of work. (b) The 4000 J of heat transfer occurs first irreversibly to a 250 K reservoir and then goes into a Carnot engine. The increase in entropy caused by the heat transfer to a colder reservoir results in a smaller work output of 2400 J. There is a permanent loss of 933 J of energy for the purpose of doing work.

When entropy increases, a certain amount of energy becomes *permanently* unavailable to do work. The energy is not lost, but its character is changed, so that some of it can never be converted to doing work—that is, to an organized force acting through a distance. For instance, in the previous example, 933 J less work was done after an increase in entropy of 9.33 J/K occurred in the 4000 J heat transfer from the 600 K reservoir to the 250 K reservoir. It can be shown that the amount of energy that becomes unavailable for work is

Equation:

$$W_{\text{unavail}} = \Delta S \cdot T_0,$$

where T_0 is the lowest temperature utilized. In the previous example,

Equation:

$$W_{\text{unavail}} = (9.33 \text{ J/K})(100 \text{ K}) = 933 \text{ J}$$

as found.

Heat Death of the Universe: An Overdose of Entropy

In the early, energetic universe, all matter and energy were easily interchangeable and identical in nature. Gravity played a vital role in the young universe. Although it may have *seemed* disorderly, and therefore, superficially entropic, in fact, there was enormous potential energy available to do work—all the future energy in the universe.

As the universe matured, temperature differences arose, which created more opportunity for work. Stars are hotter than planets, for example, which are warmer than icy asteroids, which are warmer still than the vacuum of the space between them.

Most of these are cooling down from their usually violent births, at which time they were provided with energy of their own—nuclear energy in the case of stars, volcanic energy on Earth and other planets, and so on. Without additional energy input, however, their days are numbered.

As entropy increases, less and less energy in the universe is available to do work. On Earth, we still have great stores of energy such as fossil and nuclear fuels; large-scale temperature differences, which can provide wind energy; geothermal energies due to differences in temperature in Earth's layers; and tidal energies owing to our abundance of liquid water. As these are used, a certain fraction of the energy they contain can never be converted into doing work. Eventually, all fuels will be exhausted, all temperatures will equalize, and it will be impossible for heat engines to function, or for work to be done.

Entropy increases in a closed system, such as the universe. But in parts of the universe, for instance, in the Solar system, it is not a locally closed system. Energy flows from the Sun to the planets, replenishing Earth's stores of energy. The Sun will continue to supply us with energy for about another five billion years. We will enjoy direct solar energy, as well as side effects of solar energy, such as wind power and biomass energy from photosynthetic plants. The energy from the Sun will keep our water at the liquid state, and the Moon's gravitational pull will continue to provide tidal energy. But Earth's geothermal energy will slowly run down and won't be replenished.

But in terms of the universe, and the very long-term, very large-scale picture, the entropy of the universe is increasing, and so the availability of energy to do work is constantly decreasing. Eventually, when all stars have died, all forms of potential energy have been utilized, and all temperatures have equalized (depending on the mass of the universe, either at a very high temperature following a universal contraction, or a very low one, just before all activity ceases) there will be no possibility of doing work.

Either way, the universe is destined for thermodynamic equilibrium—maximum entropy. This is often called the *heat death of the universe*, and will mean the end of all activity. However, whether the universe contracts and heats up, or continues to expand and cools down, the end is not near.

Calculations of black holes suggest that entropy can easily continue for at least 10^{100} years.

Order to Disorder

Entropy is related not only to the unavailability of energy to do work—it is also a measure of disorder. This notion was initially postulated by Ludwig Boltzmann in the 1800s. For example, melting a block of ice means taking a highly structured and orderly system of water molecules and converting it into a disorderly liquid in which molecules have no fixed positions. (See [link].) There is a large increase in entropy in the process, as seen in the following example.

Example:
Entropy Associated with Disorder
Find the increase in entropy of 1.00 kg of ice originally at 0º C that is melted to form water at 0º C.
Strategy
As before, the change in entropy can be calculated from the definition of ΔS once we find the energy Q needed to melt the ice.
Solution
The change in entropy is defined as:
Equation:

$$\mathit{\Delta}S=\frac{Q}{T}.$$

Here Q is the heat transfer necessary to melt 1.00 kg of ice and is given by
Equation:

$$Q=mL_{\text{f}},$$

where m is the mass and L_{f} is the latent heat of fusion. $L_{\text{f}}=334\text{ kJ/kg}$ for water, so that
Equation:

$$Q = (1.00\text{ kg})(334\text{ kJ/kg}) = 3.34 \times 10^5\text{ J}.$$

Now the change in entropy is positive, since heat transfer occurs into the ice to cause the phase change; thus,

Equation:

$$\Delta S = \frac{Q}{T} = \frac{3.34 \times 10^5\text{ J}}{T}.$$

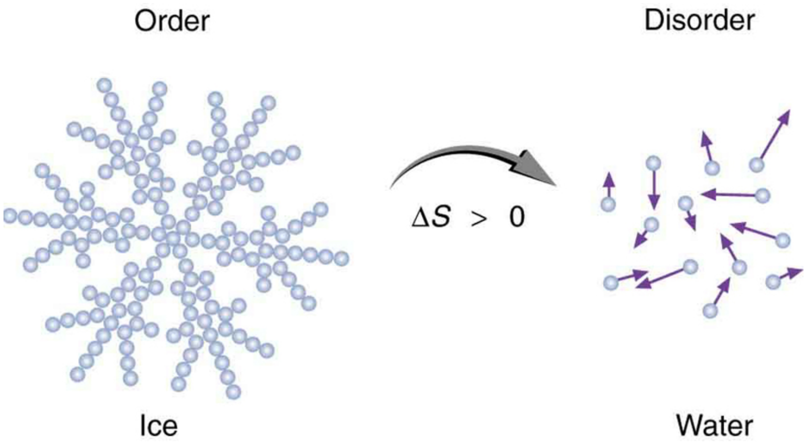
T is the melting temperature of ice. That is, $T = 0^{\text{o}}\text{C}{=}273\text{ K}$. So the change in entropy is

Equation:

$$\begin{array}{rcl}\Delta S & = & \frac{3.34{\times}10^5\text{ J}}{273\text{ K}} \\ & = & 1.22{\times}10^3\text{ J/K}.\end{array}$$

Discussion

This is a significant increase in entropy accompanying an increase in disorder.



When ice melts, it becomes more disordered and less structured. The systematic arrangement of molecules in a crystal structure is replaced by a more random and less orderly movement of molecules without fixed locations or orientations. Its entropy increases because heat transfer occurs into it. Entropy is a measure of disorder.

In another easily imagined example, suppose we mix equal masses of water originally at two different temperatures, say 20.0° C and 40.0° C. The result is water at an intermediate temperature of 30.0° C. Three outcomes have resulted: entropy has increased, some energy has become unavailable to do work, and the system has become less orderly. Let us think about each of these results.

First, entropy has increased for the same reason that it did in the example above. Mixing the two bodies of water has the same effect as heat transfer from the hot one and the same heat transfer into the cold one. The mixing decreases the entropy of the hot water but increases the entropy of the cold water by a greater amount, producing an overall increase in entropy.

Second, once the two masses of water are mixed, there is only one temperature—you cannot run a heat engine with them. The energy that could have been used to run a heat engine is now unavailable to do work.

Third, the mixture is less orderly, or to use another term, less structured. Rather than having two masses at different temperatures and with different distributions of molecular speeds, we now have a single mass with a uniform temperature.

These three results—entropy, unavailability of energy, and disorder—are not only related but are in fact essentially equivalent.

Life, Evolution, and the Second Law of Thermodynamics

Some people misunderstand the second law of thermodynamics, stated in terms of entropy, to say that the process of the evolution of life violates this law. Over time, complex organisms evolved from much simpler ancestors, representing a large decrease in entropy of the Earth’s biosphere. It is a fact that living organisms have evolved to be highly structured, and much lower in entropy than the substances from which they grow. But it is *always* possible for the entropy of one part of the universe to decrease, provided the total change in entropy of the universe increases. In equation form, we can write this as

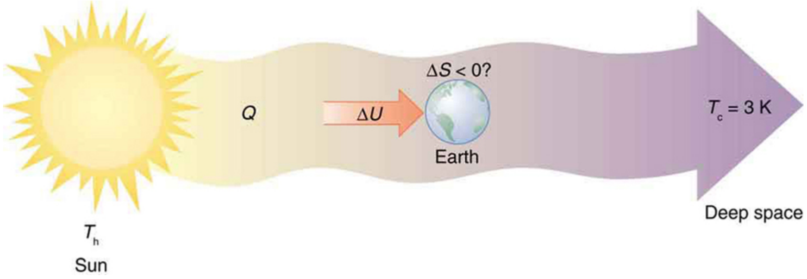
Equation:

$$\Delta S_{\text{tot}} = \Delta S_{\text{syst}} + \Delta S_{\text{envir}} > 0.$$

Thus ΔS_{syst} can be negative as long as ΔS_{envir} is positive and greater in magnitude.

How is it possible for a system to decrease its entropy? Energy transfer is necessary. If I pick up marbles that are scattered about the room and put them into a cup, my work has decreased the entropy of that system. If I gather iron ore from the ground and convert it into steel and build a bridge, my work has decreased the entropy of that system. Energy coming from the Sun can decrease the entropy of local systems on Earth—that is, ΔS_{syst} is negative. But the overall entropy of the rest of the universe increases by a greater amount—that is, ΔS_{envir} is positive and greater in magnitude. Thus, $\Delta S_{\text{tot}} = \Delta S_{\text{syst}} + \Delta S_{\text{envir}} > 0$, and the second law of thermodynamics is *not* violated.

Every time a plant stores some solar energy in the form of chemical potential energy, or an updraft of warm air lifts a soaring bird, the Earth can be viewed as a heat engine operating between a hot reservoir supplied by the Sun and a cold reservoir supplied by dark outer space—a heat engine of high complexity, causing local decreases in entropy as it uses part of the heat transfer from the Sun into deep space. There is a large total increase in entropy resulting from this massive heat transfer. A small part of this heat transfer is stored in structured systems on Earth, producing much smaller local decreases in entropy. (See [link].)



Earth’s entropy may decrease in the process of intercepting a small part of the heat transfer from the Sun into deep space. Entropy for the entire process increases greatly while Earth becomes more structured with living systems and stored energy in various forms.

Note:
PhET Explorations: Reversible Reactions
Watch a reaction proceed over time. How does total energy affect a reaction rate? Vary temperature, barrier height, and potential energies. Record concentrations and time in order to extract rate coefficients. Do temperature dependent studies to extract Arrhenius parameters. This simulation is best used with teacher guidance because it presents an analogy of chemical reactions.

Reversible Reactions

Section Summary

- Entropy is the loss of energy available to do work.
- Another form of the second law of thermodynamics states that the total entropy of a system either increases or remains constant; it never decreases.
- Entropy is zero in a reversible process; it increases in an irreversible process.
- The ultimate fate of the universe is likely to be thermodynamic equilibrium, where the universal temperature is constant and no energy is available to do work.
- Entropy is also associated with the tendency toward disorder in a closed system.

Conceptual Questions

Exercise:

Problem:

A woman shuts her summer cottage up in September and returns in June. No one has entered the cottage in the meantime. Explain what she is likely to find, in terms of the second law of thermodynamics.

Exercise:

Problem:

Consider a system with a certain energy content, from which we wish to extract as much work as possible. Should the system’s entropy be high or low? Is this orderly or disorderly? Structured or uniform? Explain briefly.

Exercise:

Problem:

Does a gas become more orderly when it liquefies? Does its entropy change? If so, does the entropy increase or decrease? Explain your answer.

Exercise:

Problem:

Explain how water’s entropy can decrease when it freezes without violating the second law of thermodynamics. Specifically, explain what happens to the entropy of its surroundings.

Exercise:

Problem:

Is a uniform-temperature gas more or less orderly than one with several different temperatures? Which is more structured? In which can heat transfer result in work done without heat transfer from another system?

Exercise:

Problem:

Give an example of a spontaneous process in which a system becomes less ordered and energy becomes less available to do work. What happens to the system’s entropy in this process?

Exercise:

Problem:

What is the change in entropy in an adiabatic process? Does this imply that adiabatic processes are reversible? Can a process be precisely adiabatic for a macroscopic system?

Exercise:

Problem:

Does the entropy of a star increase or decrease as it radiates? Does the entropy of the space into which it radiates (which has a temperature of about 3 K) increase or decrease? What does this do to the entropy of the universe?

Exercise:

Problem:

Explain why a building made of bricks has smaller entropy than the same bricks in a disorganized pile. Do this by considering the number of ways that each could be formed (the number of microstates in each macrostate).

Problem Exercises

Exercise:

Problem:

(a) On a winter day, a certain house loses $5.00\times10^8\ \text{J}$ of heat to the outside (about 500,000 Btu). What is the total change in entropy due to this heat transfer alone, assuming an average indoor temperature of $21.0^\text{o}\ \text{C}$ and an average outdoor temperature of $5.00^\text{o}\ \text{C}$? (b) This large change in entropy implies a large amount of energy has become unavailable to do work. Where do we find more energy when such energy is lost to us?

Solution:

(a) $9.78\times10^4\ \text{J/K}$

(b) In order to gain more energy, we must generate it from things within the house, like a heat pump, human bodies, and other appliances. As you know, we use a lot of energy to keep our houses warm in the winter because of the loss of heat to the outside.

Exercise:

Problem:

On a hot summer day, $4.00\times10^6\ \text{J}$ of heat transfer into a parked car takes place, increasing its temperature from $35.0^\text{o}\ \text{C}$ to $45.0^\text{o}\ \text{C}$. What is the increase in entropy of the car due to this heat transfer alone?

Exercise:

Problem:

A hot rock ejected from a volcano’s lava fountain cools from $1100^{\mathrm{o}}\ \mathrm{C}$ to $40.0^{\mathrm{o}}\ \mathrm{C}$, and its entropy decreases by 950 J/K. How much heat transfer occurs from the rock?

Solution:

$8.01 \times 10^5\ \mathrm{J}$

Exercise:

Problem:

When $1.60{\times}10^5\ \mathrm{J}$ of heat transfer occurs into a meat pie initially at $20.0^{\mathrm{o}}\ \mathrm{C}$, its entropy increases by 480 J/K. What is its final temperature?

Exercise:

Problem:

The Sun radiates energy at the rate of $3.80{\times}10^{26}\ \mathrm{W}$ from its $5500^{\mathrm{o}}\ \mathrm{C}$ surface into dark empty space (a negligible fraction radiates onto Earth and the other planets). The effective temperature of deep space is $-270^{\mathrm{o}}\ \mathrm{C}$. (a) What is the increase in entropy in one day due to this heat transfer? (b) How much work is made unavailable?

Solution:

(a) $1.04 \times 10^{31}\ \mathrm{J/K}$

(b) $3.28 \times 10^{31}\ \mathrm{J}$

Exercise:

Problem:

(a) In reaching equilibrium, how much heat transfer occurs from 1.00 kg of water at 40.0^{o} C when it is placed in contact with 1.00 kg of 20.0^{o} C water in reaching equilibrium? (b) What is the change in entropy due to this heat transfer? (c) How much work is made unavailable, taking the lowest temperature to be 20.0^{o} C? Explicitly show how you follow the steps in the Problem-Solving Strategies for Entropy.

Exercise:

Problem:

What is the decrease in entropy of 25.0 g of water that condenses on a bathroom mirror at a temperature of 35.0^{o} C, assuming no change in temperature and given the latent heat of vaporization to be 2450 kJ/kg?

Solution:

199 J/K

Exercise:

Problem:

Find the increase in entropy of 1.00 kg of liquid nitrogen that starts at its boiling temperature, boils, and warms to 20.0^{o} C at constant pressure.

Exercise:

Problem:

A large electrical power station generates 1000 MW of electricity with an efficiency of 35.0%. (a) Calculate the heat transfer to the power station, Q_{h}, in one day. (b) How much heat transfer Q_{c} occurs to the environment in one day? (c) If the heat transfer in the cooling towers is from 35.0^{o} C water into the local air mass, which increases in temperature from 18.0^{o} C to 20.0^{o} C, what is the total increase in entropy due to this heat transfer? (d) How much energy becomes unavailable to do work because of this increase in entropy, assuming an 18.0^{o} C lowest temperature? (Part of Q_{c} could be utilized to operate heat engines or for simply heating the surroundings, but it rarely is.)

Solution:

(a) $2.47\times10^{14}\text{ J}$

(b) $1.60\times10^{14}\text{ J}$

(c) $2.85\times10^{10}\text{ J/K}$

(d) $8.29\times10^{12}\text{ J}$

Exercise:

Problem:

(a) How much heat transfer occurs from 20.0 kg of $90.0^\circ\ \mathrm{C}$ water placed in contact with 20.0 kg of $10.0^\circ\ \mathrm{C}$ water, producing a final temperature of $50.0^\circ\ \mathrm{C}$? (b) How much work could a Carnot engine do with this heat transfer, assuming it operates between two reservoirs at constant temperatures of $90.0^\circ\ \mathrm{C}$ and $10.0^\circ\ \mathrm{C}$? (c) What increase in entropy is produced by mixing 20.0 kg of $90.0^\circ\ \mathrm{C}$ water with 20.0 kg of $10.0^\circ\ \mathrm{C}$ water? (d) Calculate the amount of work made unavailable by this mixing using a low temperature of $10.0^\circ\ \mathrm{C}$, and compare it with the work done by the Carnot engine. Explicitly show how you follow the steps in the Problem-Solving Strategies for Entropy. (e) Discuss how everyday processes make increasingly more energy unavailable to do work, as implied by this problem.

Glossary

entropy
: a measurement of a system's disorder and its inability to do work in a system

change in entropy
: the ratio of heat transfer to temperature Q/T

second law of thermodynamics stated in terms of entropy
: the total entropy of a system either increases or remains constant; it never decreases

Statistical Interpretation of Entropy and the Second Law of Thermodynamics: The Underlying Explanation

- Identify probabilities in entropy.
- Analyze statistical probabilities in entropic systems.

When you toss a coin a large number of times, heads and tails tend to come up in roughly equal numbers. Why doesn't heads come up 100, 90, or even 80% of the time? (credit: Jon Sullivan, PDPhoto.org)

The various ways of formulating the second law of thermodynamics tell what happens rather than why it happens. Why should heat transfer occur only from hot to cold? Why should energy become ever less available to do work? Why should the universe become increasingly disorderly? The answer is that it is a matter of overwhelming probability. Disorder is simply vastly more likely than order.

When you watch an emerging rain storm begin to wet the ground, you will notice that the drops fall in a disorganized manner both in time and in space. Some fall close together, some far apart, but they never fall in straight, orderly rows. It is not impossible for rain to fall in an orderly pattern, just highly unlikely, because there are many more disorderly ways than orderly ones. To illustrate this fact, we will examine some random processes, starting with coin tosses.

Coin Tosses

What are the possible outcomes of tossing 5 coins? Each coin can land either heads or tails. On the large scale, we are concerned only with the total heads and tails and not with the order in which heads and tails appear. The following possibilities exist:

Equation:

$$\begin{gathered}\text{5 heads, 0 tails}\\ \text{4 heads, 1 tail}\\ \text{3 heads, 2 tails}\\ \text{2 heads, 3 tails}\\ \text{1 head, 4 tails}\\ \text{0 head, 5 tails}\end{gathered}$$

These are what we call macrostates. A **macrostate** is an overall property of a system. It does not specify the details of the system, such as the order in which heads and tails occur or which coins are heads or tails.

Using this nomenclature, a system of 5 coins has the 6 possible macrostates just listed. Some macrostates are more likely to occur than others. For instance, there is only one way to get 5 heads, but there are several ways to get 3 heads and 2 tails, making the latter macrostate more probable. [link] lists of all the ways in which 5 coins can be tossed, taking into account the order in which heads and tails occur. Each sequence is called a **microstate** —a detailed description of every element of a system.

	Individual microstates	**Number of microstates**
5 heads, 0 tails	HHHHH	1
4 heads, 1 tail	HHHHT, HHHTH, HHTHH, HTHHH, THHHH	5
3 heads, 2 tails	HTHTH, THTHH, HTHHT, THHTH, THHHT HTHTH, THTHH, HTHHT, THHTH, THHHT	10
2 heads, 3 tails	TTTHH, TTHHT, THHTT, HHTTT, TTHTH, THTHT, HTHTT, THTTH, HTTHT, HTTTH	10
1 head, 4 tails	TTTTH, TTTHT, TTHTT, THTTT, HTTTT	5
0 heads, 5 tails	TTTTT	1
		Total: 32

5-Coin Toss

The macrostate of 3 heads and 2 tails can be achieved in 10 ways and is thus 10 times more probable than the one having 5 heads. Not surprisingly, it is equally probable to have the reverse, 2 heads and 3 tails. Similarly, it is equally probable to get 5 tails as it is to get 5 heads. Note that all of these conclusions are based on the crucial assumption that each microstate is equally probable. With coin tosses, this requires that the coins not be asymmetric in a way that favors one side over the other, as with loaded dice. With any system, the assumption that all microstates are equally probable must be valid, or the analysis will be erroneous.

The two most orderly possibilities are 5 heads or 5 tails. (They are more structured than the others.) They are also the least likely, only 2 out of 32 possibilities. The most disorderly possibilities are 3 heads and 2 tails and its reverse. (They are the least structured.) The most disorderly possibilities are also the most likely, with 20 out of 32 possibilities for the 3 heads and 2 tails and its reverse. If we start with an orderly array like 5 heads and toss the coins, it is very likely that we will get a less orderly array as a result, since 30 out of the 32 possibilities are less orderly. So even if you start with an orderly state, there is a strong tendency to go from order to disorder, from low entropy to high entropy. The reverse can happen, but it is unlikely.

Macrostate		**Number of microstates**
Heads	Tails	(W)
100	0	1
99	1	1.0×10^{2}
95	5	7.5×10^{7}
90	10	1.7×10^{13}

Macrostate		Number of microstates
75	25	2.4×10^{23}
60	40	1.4×10^{28}
55	45	6.1×10^{28}
51	49	9.9×10^{28}
50	50	1.0×10^{29}
49	51	9.9×10^{28}
45	55	6.1×10^{28}
40	60	1.4×10^{28}
25	75	2.4×10^{23}

Macrostate		Number of microstates
10	90	1.7×10^{13}
5	95	7.5×10^{7}
1	99	1.0×10^{2}
0	100	1
		Total: 1.27×10^{30}

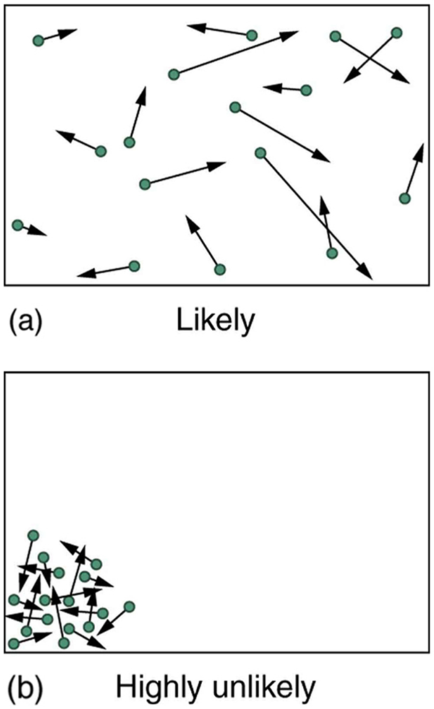
100-Coin Toss

This result becomes dramatic for larger systems. Consider what happens if you have 100 coins instead of just 5. The most orderly arrangements (most structured) are 100 heads or 100 tails. The least orderly (least structured) is that of 50 heads and 50 tails. There is only 1 way (1 microstate) to get the most orderly arrangement of 100 heads. There are 100 ways (100 microstates) to get the next most orderly arrangement of 99 heads and 1 tail (also 100 to get its reverse). And there are 1.0×10^{29} ways to get 50 heads and 50 tails, the least orderly arrangement. [link] is an abbreviated list of the various macrostates and the number of microstates for each macrostate. The total number of microstates—the total number of different ways 100 coins can be tossed—is an impressively large 1.27×10^{30}. Now, if we start with an orderly macrostate like 100 heads and toss the coins, there is a virtual certainty that we will get a less orderly macrostate. If we keep tossing the coins, it is possible, but exceedingly unlikely, that we will ever get back to the most orderly macrostate. If you tossed the coins once each second, you could expect to get either 100 heads or 100 tails once in 2×10^{22} years! This period is 1 trillion (10^{12}) times longer than the age of the universe, and so the chances are essentially zero. In contrast, there is an 8% chance of getting 50 heads, a 73% chance of getting from 45 to 55 heads, and a 96% chance of getting from 40 to 60 heads. Disorder is highly likely.

Disorder in a Gas

The fantastic growth in the odds favoring disorder that we see in going from 5 to 100 coins continues as the number of entities in the system increases. Let us now imagine applying this approach to perhaps a small sample of gas. Because counting microstates and macrostates involves statistics, this is called **statistical analysis**. The macrostates of a gas correspond to its macroscopic properties, such as volume, temperature, and pressure; and its microstates correspond to the detailed description of the positions and velocities of its atoms. Even a small amount of gas has a huge number of atoms: 1.0 cm^3 of an ideal gas at 1.0 atm and 0^{o} C has 2.7×10^{19} atoms. So each macrostate has an immense number of microstates. In plain language, this means that there are an immense number of ways in which the atoms in a gas can be arranged, while still having the same pressure, temperature, and so on.

The most likely conditions (or macrostates) for a gas are those we see all the time—a random distribution of atoms in space with a Maxwell-Boltzmann distribution of speeds in random directions, as predicted by kinetic theory. This is the most disorderly and least structured condition we can imagine. In contrast, one type of very orderly and structured macrostate has all of the atoms in one corner of a container with identical velocities. There are very few ways to accomplish this (very few microstates corresponding to it), and so it is exceedingly unlikely ever to occur. (See [link](b).) Indeed, it is so unlikely that we have a law saying that it is impossible, which has never been observed to be violated—the second law of thermodynamics.



(a) The ordinary state of gas in a container is a disorderly, random distribution of atoms or molecules with a Maxwell-Boltzmann distribution of speeds. It is so unlikely that these atoms or molecules would ever end up in one corner of the container that it might as well be impossible. (b) With energy transfer, the gas can be forced into one corner and its entropy greatly reduced. But left alone, it will spontaneously increase its entropy and return to the normal conditions, because they are immensely more likely.

The disordered condition is one of high entropy, and the ordered one has low entropy. With a transfer of energy from another system, we could force all of the atoms into one corner and have a local decrease in entropy, but at the cost of an overall increase in entropy of the universe. If the atoms start out in one corner, they will quickly disperse and become uniformly distributed and will never return to the orderly original state ([link](b)). Entropy will increase. With such a large sample of atoms, it is possible—but unimaginably unlikely—for entropy to decrease. Disorder is vastly more likely than order.

The arguments that disorder and high entropy are the most probable states are quite convincing. The great Austrian physicist Ludwig Boltzmann (1844–1906)—who, along with Maxwell, made so many contributions to kinetic theory—proved that the entropy of a system in a given state (a macrostate) can be written as

Equation:

$$S = k \ln W,$$

where $k = 1.38 \times 10^{-23}\ \text{J/K}$ is Boltzmann’s constant, and $\ln W$ is the natural logarithm of the number of microstates W corresponding to the given macrostate. W is proportional to the probability that the macrostate will occur. Thus entropy is directly related to the probability of a state—the more likely the state, the greater its entropy. Boltzmann proved that this expression for S is equivalent to the definition $\Delta S = Q/T$, which we have used extensively.

Thus the second law of thermodynamics is explained on a very basic level: entropy either remains the same or increases in every process. This phenomenon is due to the extraordinarily small probability of a decrease, based on the extraordinarily larger number of microstates in systems with greater entropy. Entropy *can* decrease, but for any macroscopic system, this outcome is so unlikely that it will never be observed.

Example:

Entropy Increases in a Coin Toss

Suppose you toss 100 coins starting with 60 heads and 40 tails, and you get the most likely result, 50 heads and 50 tails. What is the change in entropy?

Strategy

Noting that the number of microstates is labeled W in [link] for the 100-coin toss, we can use $\Delta S = S_{\text{f}} - S_{\text{i}} = k\text{ln}W_{\text{f}} - k\text{ln}W_{\text{i}}$ to calculate the change in entropy.

Solution

The change in entropy is

Equation:

$$\Delta S = S_{\text{f}} – S_{\text{i}} = k\text{ln}W_{\text{f}} – k\text{ln}W_{\text{i,}}$$

where the subscript i stands for the initial 60 heads and 40 tails state, and the subscript f for the final 50 heads and 50 tails state. Substituting the values for W from [link] gives

Equation:

$$\begin{array}{lll}\Delta S & = & (1.38 \times 10^{–23}\text{ J/K})[\text{ln}(1.0 \times 10^{29}) – \text{ln}(1.4 \times 10^{28})] \\ & = & 2.7 \times 10^{–23}\text{ J/K}\end{array}$$

Discussion

This increase in entropy means we have moved to a less orderly situation. It is not impossible for further tosses to produce the initial state of 60 heads and 40 tails, but it is less likely. There is about a 1 in 90 chance for that decrease in entropy ($– 2.7 \times 10^{–23}\text{ J/K}$) to occur. If we calculate the decrease in entropy to move to the most orderly state, we get

$\Delta S = -92 \times 10^{-23}\ \text{J/K}$. There is about a 1 in 10^{30} chance of this change occurring. So while very small decreases in entropy are unlikely, slightly greater decreases are impossibly unlikely. These probabilities imply, again, that for a macroscopic system, a decrease in entropy is impossible. For example, for heat transfer to occur spontaneously from 1.00 kg of 0^{o}C ice to its 0^{o}C environment, there would be a decrease in entropy of $1.22 \times 10^{3}\ \text{J/K}$. Given that a ΔS of $10^{-21}\ \text{J/K}$ corresponds to about a 1 in 10^{30} chance, a decrease of this size ($10^{3}\ \text{J/K}$) is an *utter* impossibility. Even for a milligram of melted ice to spontaneously refreeze is impossible.

Note:

Problem-Solving Strategies for Entropy

1. *Examine the situation to determine if entropy is involved.*
2. *Identify the system of interest and draw a labeled diagram of the system showing energy flow.*
3. *Identify exactly what needs to be determined in the problem (identify the unknowns).* A written list is useful.
4. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* You must carefully identify the heat transfer, if any, and the temperature at which the process takes place. It is also important to identify the initial and final states.
5. *Solve the appropriate equation for the quantity to be determined (the unknown).* Note that the change in entropy can be determined between any states by calculating it for a reversible process.
6. *Substitute the known value along with their units into the appropriate equation, and obtain numerical solutions complete with units.*
7. *To see if it is reasonable: Does it make sense?* For example, total entropy should increase for any real process or be constant for a reversible process. Disordered states should be more probable and have greater entropy than ordered states.

Section Summary

- Disorder is far more likely than order, which can be seen statistically.
- The entropy of a system in a given state (a macrostate) can be written as

  Equation:

  $$S = k \ln W,$$

  where $k = 1.38 \times 10^{-23}\ \text{J/K}$ is Boltzmann's constant, and $\ln W$ is the natural logarithm of the number of microstates W corresponding to the given macrostate.

Conceptual Questions

Exercise:

Problem:

Explain why a building made of bricks has smaller entropy than the same bricks in a disorganized pile. Do this by considering the number of ways that each could be formed (the number of microstates in each macrostate).

Problem Exercises

Exercise:

Problem:

Using [link], verify the contention that if you toss 100 coins each second, you can expect to get 100 heads or 100 tails once in 2×10^{22} years; calculate the time to two-digit accuracy.

Solution:

It should happen twice in every 1.27×10^{30} s or once in every

$$6.35 \times 10^{29}\ \text{s} = \left(6.35 \times 10^{29}\ \text{s}\right)\left(\frac{1\ \text{h}}{3600\ \text{s}}\right)\left(\frac{1\ \text{d}}{24\ \text{h}}\right)\left(\frac{1\ \text{y}}{365.25\ \text{d}}\right) = 2.0 \times 10^{22}\ \text{y}$$

Exercise:

Problem:

What percent of the time will you get something in the range from 60 heads and 40 tails through 40 heads and 60 tails when tossing 100 coins? The total number of microstates in that range is 1.22×10^{30}. (Consult [link].)

Exercise:

Problem:

(a) If tossing 100 coins, how many ways (microstates) are there to get the three most likely macrostates of 49 heads and 51 tails, 50 heads and 50 tails, and 51 heads and 49 tails? (b) What percent of the total possibilities is this? (Consult [link].)

Solution:

(a) 3.0×10^{29}

(b) 24%

Exercise:

Problem:

(a) What is the change in entropy if you start with 100 coins in the 45 heads and 55 tails macrostate, toss them, and get 51 heads and 49 tails? (b) What if you get 75 heads and 25 tails? (c) How much more likely is 51 heads and 49 tails than 75 heads and 25 tails? (d) Does either outcome violate the second law of thermodynamics?

Exercise:

Problem:

(a) What is the change in entropy if you start with 10 coins in the 5 heads and 5 tails macrostate, toss them, and get 2 heads and 8 tails? (b) How much more likely is 5 heads and 5 tails than 2 heads and 8 tails? (Take the ratio of the number of microstates to find out.) (c) If you were betting on 2 heads and 8 tails would you accept odds of 252 to 45? Explain why or why not.

Solution:

(a) $-2.38\times10^{-23}\ \text{J/K}$

(b) 5.6 times more likely

(c) If you were betting on two heads and 8 tails, the odds of breaking even are 252 to 45, so on average you would break even. So, no, you wouldn’t bet on odds of 252 to 45.

Macrostate		**Number of Microstates (*W*)**
Heads	Tails	
10	0	1
9	1	10
8	2	45
7	3	120

Macrostate		**Number of Microstates (*W*)**
6	4	210
5	5	252
4	6	210
3	7	120
2	8	45
1	9	10
0	10	1
		Total: 1024

10-Coin Toss

Exercise:

Problem:

(a) If you toss 10 coins, what percent of the time will you get the three most likely macrostates (6 heads and 4 tails, 5 heads and 5 tails, 4 heads and 6 tails)? (b) You can realistically toss 10 coins and count the number of heads and tails about twice a minute. At that rate, how long will it take on average to get either 10 heads and 0 tails or 0 heads and 10 tails?

Exercise:

Problem:

(a) Construct a table showing the macrostates and all of the individual microstates for tossing 6 coins. (Use [link] as a guide.) (b) How many macrostates are there? (c) What is the total number of microstates? (d) What percent chance is there of tossing 5 heads and 1 tail? (e) How much more likely are you to toss 3 heads and 3 tails than 5 heads and 1 tail? (Take the ratio of the number of microstates to find out.)

Solution:

(b) 7

(c) 64

(d) 9.38%

(e) 3.33 times more likely (20 to 6)

Exercise:

Problem:

In an air conditioner, 12.65 MJ of heat transfer occurs from a cold environment in 1.00 h. (a) What mass of ice melting would involve the same heat transfer? (b) How many hours of operation would be equivalent to melting 900 kg of ice? (c) If ice costs 20 cents per kg, do you think the air conditioner could be operated more cheaply than by simply using ice? Describe in detail how you evaluate the relative costs.

Glossary

macrostate
: an overall property of a system

microstate
: each sequence within a larger macrostate

statistical analysis

using statistics to examine data, such as counting microstates and macrostates

Introduction to Oscillatory Motion and Waves class="introduction"

There are at least four types of waves in this picture—only the water waves are evident. There are also sound waves, light waves, and waves on the guitar strings. (credit: John Norton)

What do an ocean buoy, a child in a swing, the cone inside a speaker, a guitar, atoms in a crystal, the motion of chest cavities, and the beating of hearts all have in common? They all **oscillate**—-that is, they move back and forth between two points. Many systems oscillate, and they have certain characteristics in common. All oscillations involve force and energy. You push a child in a swing to get the motion started. The energy of atoms vibrating in a crystal can be increased with heat. You put energy into a guitar string when you pluck it.

Some oscillations create **waves**. A guitar creates sound waves. You can make water waves in a swimming pool by slapping the water with your hand. You can no doubt think of other types of waves. Some, such as water waves, are visible. Some, such as sound waves, are not. But *every wave is a disturbance that moves from its source and carries energy*. Other examples of waves include earthquakes and visible light. Even subatomic particles, such as electrons, can behave like waves.

By studying oscillatory motion and waves, we shall find that a small number of underlying principles describe all of them and that wave phenomena are more common than you have ever imagined. We begin by studying the type of force that underlies the simplest oscillations and waves. We will then expand our exploration of oscillatory motion and waves to include concepts such as simple harmonic motion, uniform circular motion, and damped harmonic motion. Finally, we will explore what happens when two or more waves share the same space, in the phenomena known as superposition and interference.

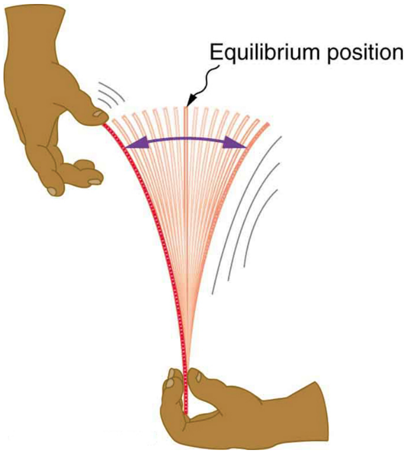
Glossary

oscillate
: moving back and forth regularly between two points

wave
: a disturbance that moves from its source and carries energy

Hooke’s Law: Stress and Strain Revisited

- Explain Newton’s third law of motion with respect to stress and deformation.
- Describe the restoration of force and displacement.
- Calculate the energy in Hooke’s Law of deformation, and the stored energy in a spring.



When displaced from its vertical equilibrium position, this plastic ruler oscillates back and forth because of the restoring force opposing displacement. When the ruler is on the left, there is a force to the right, and vice versa.

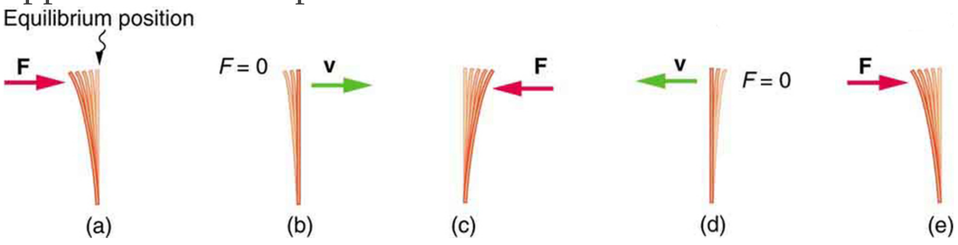
Newton’s first law implies that an object oscillating back and forth is experiencing forces. Without force, the object would move in a straight line at a constant speed rather than oscillate. Consider, for example, plucking a plastic ruler to the left as shown in [link]. The deformation of the ruler creates a force in the opposite direction, known as a **restoring force**. Once released, the restoring force causes the ruler to move back toward its stable equilibrium position, where the net force on it is zero. However, by the time the ruler gets there, it gains momentum and continues to move to the right, producing the opposite deformation. It is then forced to the left, back through equilibrium, and the process is repeated until dissipative forces dampen the motion. These forces remove mechanical energy from the system, gradually reducing the motion until the ruler comes to rest.

The simplest oscillations occur when the restoring force is directly proportional to displacement. When stress and strain were covered in Newton’s Third Law of Motion, the name was given to this relationship between force and displacement was Hooke’s law:

Equation:

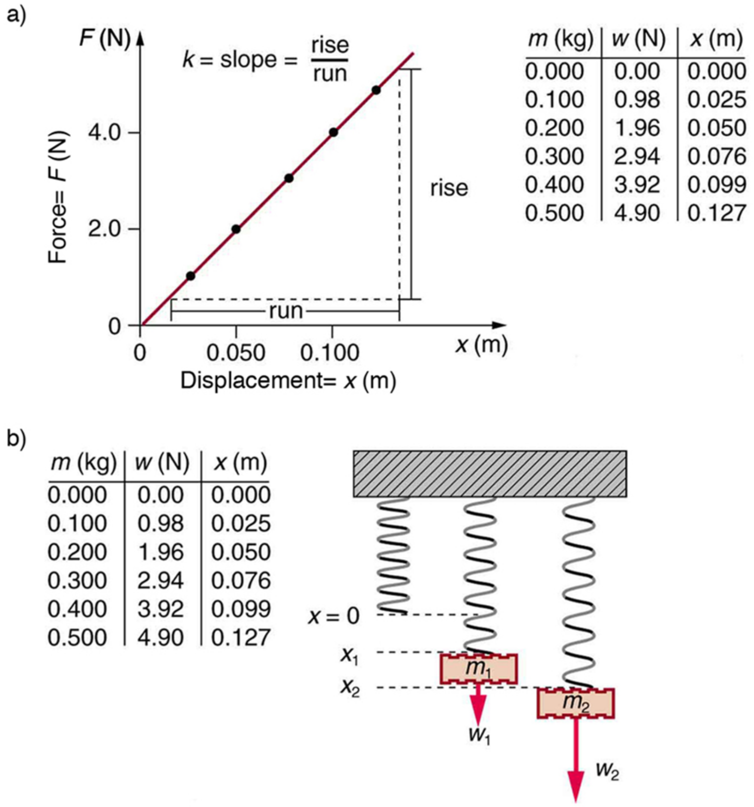
$$F=-\text{kx}.$$

Here, F is the restoring force, x is the displacement from equilibrium or **deformation**, and k is a constant related to the difficulty in deforming the system. The minus sign indicates the restoring force is in the direction opposite to the displacement.



(a) The plastic ruler has been released, and the restoring force is returning the ruler to its equilibrium position. (b) The net force is zero at the equilibrium position, but the ruler has momentum and continues to move to the right. (c) The restoring force is in the opposite direction. It stops the ruler and moves it back toward equilibrium again. (d) Now the ruler has momentum to the left. (e) In the absence of damping (caused by frictional forces), the ruler reaches its original position. From there, the motion will repeat itself.

The **force constant** k is related to the rigidity (or stiffness) of a system—the larger the force constant, the greater the restoring force, and the stiffer the system. The units of k are newtons per meter (N/m). For example, k is directly related to Young’s modulus when we stretch a string. [link] shows a graph of the absolute value of the restoring force versus the displacement for a system that can be described by Hooke’s law—a simple spring in this case. The slope of the graph equals the force constant k in newtons per meter. A common physics laboratory exercise is to measure restoring forces created by springs, determine if they follow Hooke’s law, and calculate their force constants if they do.



m (kg)	w (N)	x (m)
0.000	0.00	0.000
0.100	0.98	0.025
0.200	1.96	0.050
0.300	2.94	0.076
0.400	3.92	0.099
0.500	4.90	0.127

m (kg)	w (N)	x (m)
0.000	0.00	0.000
0.100	0.98	0.025
0.200	1.96	0.050
0.300	2.94	0.076
0.400	3.92	0.099
0.500	4.90	0.127

(a) A graph of absolute value of the restoring force versus displacement is displayed. The fact that the graph is a straight line means that the system obeys Hooke’s law. The slope of the graph is the force constant k. (b) The data in the graph were generated by measuring the displacement of a spring from equilibrium while supporting various weights. The restoring force equals the weight supported, if the mass is stationary.

Example:
How Stiff Are Car Springs?

The mass of a car increases due to the introduction of a passenger. This affects the displacement of the car on its suspension system. (credit: exfordy on Flickr)

What is the force constant for the suspension system of a car that settles 1.20 cm when an 80.0-kg person gets in?

Strategy

Consider the car to be in its equilibrium position $x = 0$ before the person gets in. The car then settles down 1.20 cm, which means it is displaced to a position $x = -1.20 \times 10^{-2}$ m. At that point, the springs supply a restoring force F equal to the person's weight $w = \text{mg} = (80.0\text{ kg})\left(9.80\text{ m/s}^2\right) = 784\text{ N}$. We take this force to be F in Hooke's law. Knowing F and x, we can then solve the force constant k.

Solution

1. Solve Hooke's law, $F = -\text{kx}$, for k:
   Equation:

$$k = -\frac{F}{x}.$$

   Substitute known values and solve k:
   Equation:

$$\begin{array}{lll} k & = & -\frac{784\text{ N}}{-1.20\times10^{-2}\text{ m}} \\ & = & 6.53 \times 10^4\text{ N/m.} \end{array}$$

Discussion

Note that F and x have opposite signs because they are in opposite directions—the restoring force is up, and the displacement is down. Also, note that the car would oscillate up and down when the person got in if it were not for damping (due to frictional forces) provided by shock absorbers. Bouncing cars are a sure sign of bad shock absorbers.

Energy in Hooke’s Law of Deformation

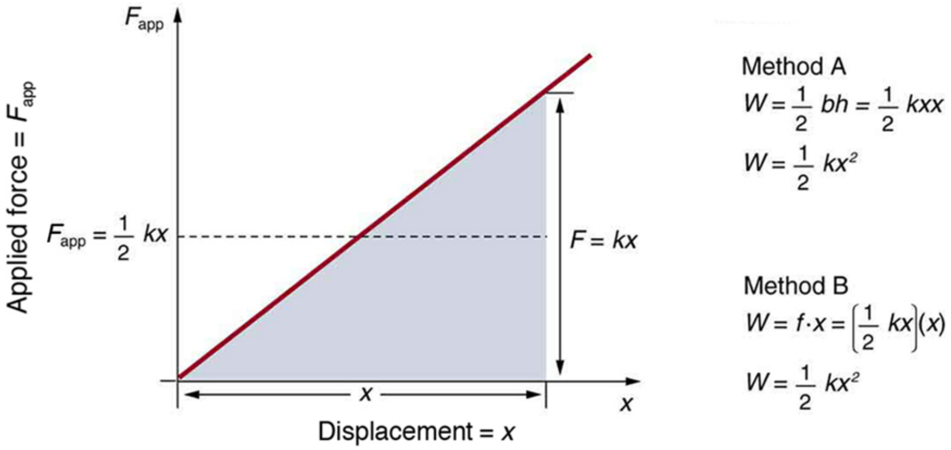
In order to produce a deformation, work must be done. That is, a force must be exerted through a distance, whether you pluck a guitar string or compress a car spring. If the only result is deformation, and no work goes into thermal, sound, or kinetic energy, then all the work is initially stored in the deformed object as some form of potential energy. The potential energy stored in a spring is $\text{PE}_{\text{el}} = \frac{1}{2}\text{kx}^2$. Here, we generalize the idea to elastic potential energy for a deformation of any system that can be described by Hooke’s law. Hence,

Equation:

$$\text{PE}_{\text{el}} = \frac{1}{2}\text{kx}^2,$$

where PE_{el} is the **elastic potential energy** stored in any deformed system that obeys Hooke’s law and has a displacement x from equilibrium and a force constant k.

It is possible to find the work done in deforming a system in order to find the energy stored. This work is performed by an applied force F_{app}. The applied force is exactly opposite to the restoring force (action-reaction), and so $F_{\text{app}} = \text{kx}$. [link] shows a graph of the applied force versus deformation x for a system that can be described by Hooke’s law. Work done on the system is force multiplied by distance, which equals the area under the curve or $(1/2)\text{kx}^2$(Method A in the figure). Another way to determine the work is to note that the force increases linearly from 0 to kx, so that the average force is $(1/2)$ kx, the distance moved is x, and thus $W = F_{\text{app}}d = [(1/2)\text{kx}](x) = (1/2)\text{kx}^2$ (Method B in the figure).

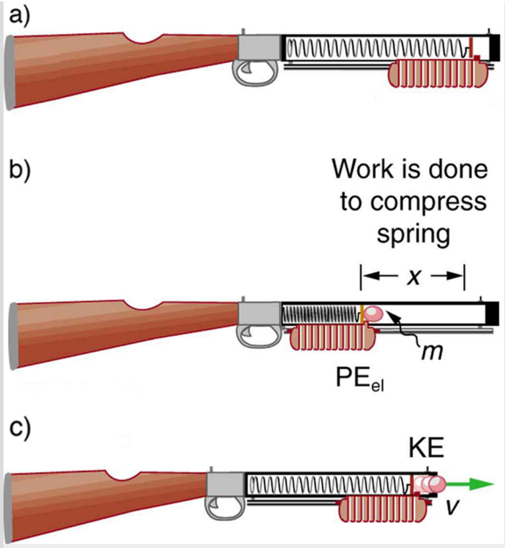


A graph of applied force versus distance for the deformation of a system that can be described by Hooke's law is displayed. The work done on the system equals the area under the graph or the area of the triangle, which is half its base multiplied by its height, or $W = (1/2)\mathrm{kx}^2$.

Example:

Calculating Stored Energy: A Tranquilizer Gun Spring

We can use a toy gun's spring mechanism to ask and answer two simple questions: (a) How much energy is stored in the spring of a tranquilizer gun that has a force constant of 50.0 N/m and is compressed 0.150 m? (b) If you neglect friction and the mass of the spring, at what speed will a 2.00-g projectile be ejected from the gun?



(a) In this image of the gun, the spring is uncompressed before being cocked. (b) The spring has been compressed a distance x, and the projectile is in place. (c) When released, the spring converts elastic potential energy PE_{el} into kinetic energy.

Strategy for a

(a): The energy stored in the spring can be found directly from elastic potential energy equation, because k and x are given.

Solution for a

Entering the given values for k and x yields

Equation:

$$\begin{aligned}\text{PE}_{\text{el}} &= \frac{1}{2}\text{kx}^2 = \frac{1}{2}(50.0\text{ N/m})(0.150\text{ m})^2 = 0.563\text{ N}\cdot\text{m} \\ &= 0.563\text{ J}\end{aligned}$$

Strategy for b

Because there is no friction, the potential energy is converted entirely into kinetic energy. The expression for kinetic energy can be solved for the projectile's speed.

Solution for b

1. Identify known quantities:
   Equation:

$$\text{KE}_\text{f} = \text{PE}_\text{el} \text{ or } 1/2mv^2 = (1/2)kx^2 = \text{PE}_\text{el} = 0.563 \text{ J}$$

2. Solve for v:
   Equation:

$$v = \left[\frac{2\text{PE}_\text{el}}{m}\right]^{1/2} = \left[\frac{2(0.563 \text{ J})}{0.002 \text{ kg}}\right]^{1/2} = 23.7(\text{J/kg})^{1/2}$$

3. Convert units: 23.7 m/s

Discussion

(a) and (b): This projectile speed is impressive for a tranquilizer gun (more than 80 km/h). The numbers in this problem seem reasonable. The force needed to compress the spring is small enough for an adult to manage, and the energy imparted to the dart is small enough to limit the damage it might do. Yet, the speed of the dart is great enough for it to travel an acceptable distance.

Exercise:
Check your Understanding

Problem:

Envision holding the end of a ruler with one hand and deforming it with the other. When you let go, you can see the oscillations of the ruler. In what way could you modify this simple experiment to increase the rigidity of the system?

Solution:
Answer

You could hold the ruler at its midpoint so that the part of the ruler that oscillates is half as long as in the original experiment.

Exercise:
Check your Understanding

Problem:

If you apply a deforming force on an object and let it come to equilibrium, what happened to the work you did on the system?

Solution:
Answer

It was stored in the object as potential energy.

Section Summary

- An oscillation is a back and forth motion of an object between two points of deformation.
- An oscillation may create a wave, which is a disturbance that propagates from where it was created.
- The simplest type of oscillations and waves are related to systems that can be described by Hooke's law:
  Equation:

  $$F=-\text{kx},$$

  where F is the restoring force, x is the displacement from equilibrium or deformation, and k is the force constant of the system.
- Elastic potential energy PE_{el} stored in the deformation of a system that can be described by Hooke's law is given by
  Equation:

$$\mathrm{PE_{el}} = (1/2)kx^2.$$

Conceptual Questions

Exercise:

Problem:

Describe a system in which elastic potential energy is stored.

Problems & Exercises

Exercise:

Problem:

Fish are hung on a spring scale to determine their mass (most fishermen feel no obligation to truthfully report the mass).

(a) What is the force constant of the spring in such a scale if it the spring stretches 8.00 cm for a 10.0 kg load?

(b) What is the mass of a fish that stretches the spring 5.50 cm?

(c) How far apart are the half-kilogram marks on the scale?

Solution:

(a) $1.23 \times 10^3 \text{ N/m}$

(b) 6.88 kg

(c) 4.00 mm

Exercise:

Problem:

It is weigh-in time for the local under-85-kg rugby team. The bathroom scale used to assess eligibility can be described by Hooke's law and is depressed 0.75 cm by its maximum load of 120 kg. (a) What is the spring's effective spring constant? (b) A player stands on the scales and depresses it by 0.48 cm. Is he eligible to play on this under-85 kg team?

Exercise:

Problem:

One type of BB gun uses a spring-driven plunger to blow the BB from its barrel. (a) Calculate the force constant of its plunger's spring if you must compress it 0.150 m to drive the 0.0500-kg plunger to a top speed of 20.0 m/s. (b) What force must be exerted to compress the spring?

Solution:

(a) 889 N/m

(b) 133 N

Exercise:

Problem:

(a) The springs of a pickup truck act like a single spring with a force constant of $1.30 \times 10^5\ \text{N/m}$. By how much will the truck be depressed by its maximum load of 1000 kg?

(b) If the pickup truck has four identical springs, what is the force constant of each?

Exercise:

Problem:

When an 80.0-kg man stands on a pogo stick, the spring is compressed 0.120 m.

(a) What is the force constant of the spring? (b) Will the spring be compressed more when he hops down the road?

Solution:

(a) $6.53 \times 10^3 \text{ N/m}$

(b) Yes

Exercise:

Problem:

A spring has a length of 0.200 m when a 0.300-kg mass hangs from it, and a length of 0.750 m when a 1.95-kg mass hangs from it. (a) What is the force constant of the spring? (b) What is the unloaded length of the spring?

Glossary

deformation
: displacement from equilibrium

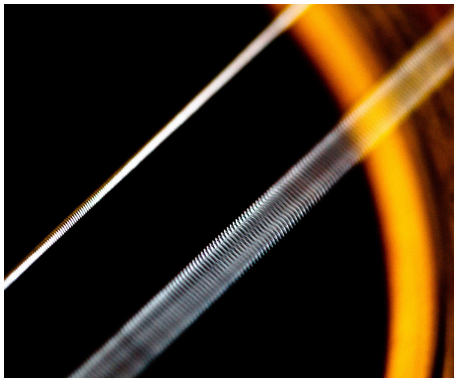
elastic potential energy
: potential energy stored as a result of deformation of an elastic object, such as the stretching of a spring

force constant
: a constant related to the rigidity of a system: the larger the force constant, the more rigid the system; the force constant is represented by *k*

restoring force
: force acting in opposition to the force caused by a deformation

Period and Frequency in Oscillations

- Observe the vibrations of a guitar string.
- Determine the frequency of oscillations.



The strings on this guitar vibrate at regular time intervals. (credit: JAR)

When you pluck a guitar string, the resulting sound has a steady tone and lasts a long time. Each successive vibration of the string takes the same time as the previous one. We define **periodic motion** to be a motion that repeats itself at regular time intervals, such as exhibited by the guitar string or by an object on a spring moving up and down. The time to complete one oscillation remains constant and is called the **period** T. Its units are usually seconds, but may be any convenient unit of time. The word period refers to the time for some event whether repetitive or not; but we shall be primarily interested in periodic motion, which is by definition repetitive. A concept closely related to period is the frequency of an event. For example, if you get a paycheck twice a month, the frequency of payment is two per month and the period between checks is half a month. **Frequency** f is defined to be the number of events per unit time. For periodic motion, frequency is the number of oscillations per unit time. The relationship between frequency and period is

Equation:

$$f=\frac{1}{T}.$$

The SI unit for frequency is the *cycle per second*, which is defined to be a *hertz* (Hz):

Equation:

$$1\text{ Hz}=1\frac{\text{cycle}}{\text{sec}}\text{or }1\text{ Hz}=\frac{1}{\text{s}}$$

A cycle is one complete oscillation. Note that a vibration can be a single or multiple event, whereas oscillations are usually repetitive for a significant number of cycles.

Example:

Determine the Frequency of Two Oscillations: Medical Ultrasound and the Period of Middle C

We can use the formulas presented in this module to determine both the frequency based on known oscillations and the oscillation based on a known frequency. Let's try one example of each. (a) A medical imaging device produces ultrasound by oscillating with a period of 0.400 µs. What is the frequency of this oscillation? (b) The frequency of middle C on a typical musical instrument is 264 Hz. What is the time for one complete oscillation?

Strategy

Both questions (a) and (b) can be answered using the relationship between period and frequency. In question (a), the period T is given and we are asked to find frequency f. In question (b), the frequency f is given and we are asked to find the period T.

Solution a

1. Substitute 0.400 µs for T in $f=\frac{1}{T}$:

   Equation:

$$f = \frac{1}{T} = \frac{1}{0.400 \times 10^{-6}\ \text{s}}.$$

Solve to find
Equation:

$$f = 2.50 \times 10^{6}\ \text{Hz}.$$

Discussion a
The frequency of sound found in (a) is much higher than the highest frequency that humans can hear and, therefore, is called ultrasound. Appropriate oscillations at this frequency generate ultrasound used for noninvasive medical diagnoses, such as observations of a fetus in the womb.
Solution b

1. Identify the known values:

   The time for one complete oscillation is the period T:
   Equation:

$$f = \frac{1}{T}.$$

2. Solve for T:
   Equation:

$$T = \frac{1}{f}.$$

3. Substitute the given value for the frequency into the resulting expression:
   Equation:

$$T = \frac{1}{f} = \frac{1}{264\ \text{Hz}} = \frac{1}{264\ \text{cycles/s}} = 3.79 \times 10^{-3}\ \text{s} = 3.79\ \text{ms}.$$

Discussion
The period found in (b) is the time per cycle, but this value is often quoted as simply the time in convenient units (ms or milliseconds in this case).

Exercise:
Check your Understanding

Problem:

Identify an event in your life (such as receiving a paycheck) that occurs regularly. Identify both the period and frequency of this event.

Solution:

I visit my parents for dinner every other Sunday. The frequency of my visits is 26 per calendar year. The period is two weeks.

Section Summary

- Periodic motion is a repetitious oscillation.
- The time for one oscillation is the period T.
- The number of oscillations per unit time is the frequency f.
- These quantities are related by
  Equation:

$$f = \frac{1}{T}.$$

Problems & Exercises

Exercise:

Problem: What is the period of 60.0 Hz electrical power?

Solution:

16.7 ms

Exercise:

Problem:

If your heart rate is 150 beats per minute during strenuous exercise, what is the time per beat in units of seconds?

Solution:

$0.400\ \text{s/beats}$

Exercise:

Problem:

Find the frequency of a tuning fork that takes 2.50×10^{-3} s to complete one oscillation.

Solution:

400 Hz

Exercise:

Problem:

A stroboscope is set to flash every 8.00×10^{-5} s. What is the frequency of the flashes?

Solution:

12,500 Hz

Exercise:

Problem:

A tire has a tread pattern with a crevice every 2.00 cm. Each crevice makes a single vibration as the tire moves. What is the frequency of these vibrations if the car moves at 30.0 m/s?

Solution:

1.50 kHz

Exercise:

Problem: **Engineering Application**

Each piston of an engine makes a sharp sound every other revolution of the engine. (a) How fast is a race car going if its eight-cylinder engine emits a sound of frequency 750 Hz, given that the engine makes 2000 revolutions per kilometer? (b) At how many revolutions per minute is the engine rotating?

Solution:

(a) 93.8 m/s

(b) $11.3 \times 10^3 \text{ rev/min}$

Glossary

period
time it takes to complete one oscillation

periodic motion
motion that repeats itself at regular time intervals

frequency
number of events per unit of time

Simple Harmonic Motion: A Special Periodic Motion

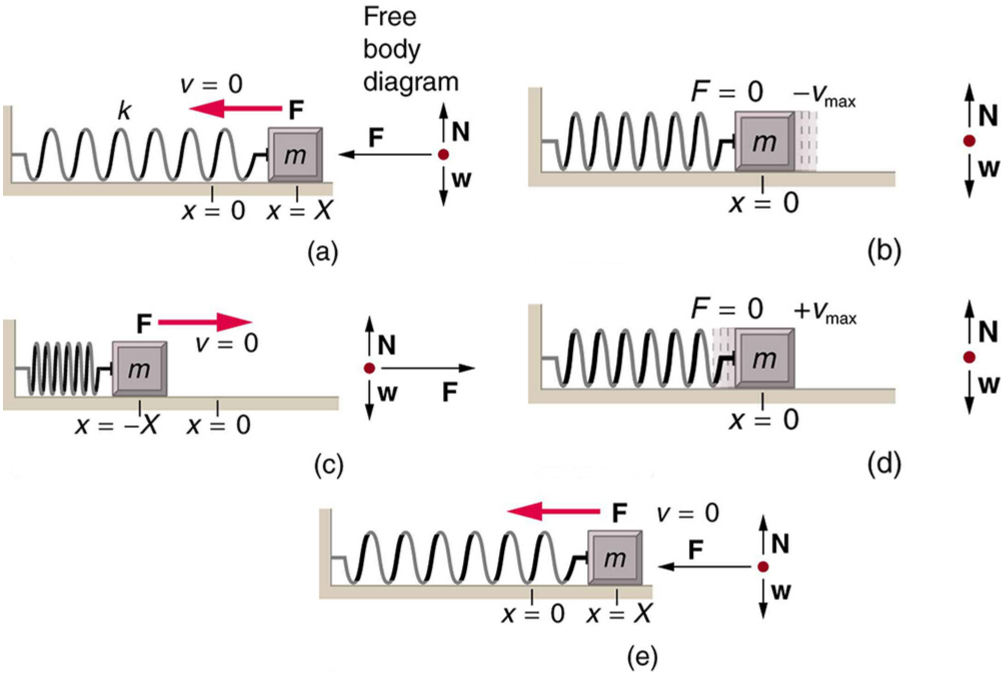
- Describe a simple harmonic oscillator.
- Explain the link between simple harmonic motion and waves.

The oscillations of a system in which the net force can be described by Hooke’s law are of special importance, because they are very common. They are also the simplest oscillatory systems. **Simple Harmonic Motion** (SHM) is the name given to oscillatory motion for a system where the net force can be described by Hooke’s law, and such a system is called a **simple harmonic oscillator**. If the net force can be described by Hooke’s law and there is no *damping* (by friction or other non-conservative forces), then a simple harmonic oscillator will oscillate with equal displacement on either side of the equilibrium position, as shown for an object on a spring in [link]. The maximum displacement from equilibrium is called the **amplitude** X. The units for amplitude and displacement are the same, but depend on the type of oscillation. For the object on the spring, the units of amplitude and displacement are meters; whereas for sound oscillations, they have units of pressure (and other types of oscillations have yet other units). Because amplitude is the maximum displacement, it is related to the energy in the oscillation.

Note:

Take-Home Experiment: SHM and the Marble

Find a bowl or basin that is shaped like a hemisphere on the inside. Place a marble inside the bowl and tilt the bowl periodically so the marble rolls from the bottom of the bowl to equally high points on the sides of the bowl. Get a feel for the force required to maintain this periodic motion. What is the restoring force and what role does the force you apply play in the simple harmonic motion (SHM) of the marble?



An object attached to a spring sliding on a frictionless surface is an uncomplicated simple harmonic oscillator. When displaced from equilibrium, the object performs simple harmonic motion that has an amplitude X and a period T. The object’s maximum speed occurs as it passes through equilibrium. The stiffer the spring is, the smaller the period T. The greater the mass of the object is, the greater the period T.

What is so significant about simple harmonic motion? One special thing is that the period T and frequency f of a simple harmonic oscillator are independent of amplitude. The string of a guitar, for example, will oscillate with the same frequency whether plucked gently or hard. Because the period is constant, a simple harmonic oscillator can be used as a clock.

Two important factors do affect the period of a simple harmonic oscillator. The period is related to how stiff the system is. A very stiff object has a large force constant k, which causes the system to have a smaller period. For example, you can adjust a diving board’s stiffness—the stiffer it is, the faster it vibrates, and the shorter its period. Period also depends on the mass of the oscillating system. The more massive the system is, the longer the period. For example, a heavy person on a diving board bounces up and down more slowly than a light one.

In fact, the mass m and the force constant k are the *only* factors that affect the period and frequency of simple harmonic motion.

Note:
Period of Simple Harmonic Oscillator
The *period of a simple harmonic oscillator* is given by
Equation:

$$T = 2\pi\sqrt{\frac{m}{k}}$$

and, because $f = 1/T$, the *frequency of a simple harmonic oscillator* is
Equation:

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}.$$

Note that neither T nor f has any dependence on amplitude.

Note:
Take-Home Experiment: Mass and Ruler Oscillations
Find two identical wooden or plastic rulers. Tape one end of each ruler firmly to the edge of a table so that the length of each ruler that protrudes from the table is the same. On the free end of one ruler tape a heavy object such as a few large coins. Pluck the ends of the rulers at the same time and observe which one undergoes more cycles in a time period, and measure the period of oscillation of each of the rulers.

Example:
Calculate the Frequency and Period of Oscillations: Bad Shock Absorbers in a Car

If the shock absorbers in a car go bad, then the car will oscillate at the least provocation, such as when going over bumps in the road and after stopping (See [link]). Calculate the frequency and period of these oscillations for such a car if the car’s mass (including its load) is 900 kg and the force constant (k) of the suspension system is $6.53 \times 10^4\ \text{N/m}$.

Strategy

The frequency of the car’s oscillations will be that of a simple harmonic oscillator as given in the equation $f = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$. The mass and the force constant are both given.

Solution

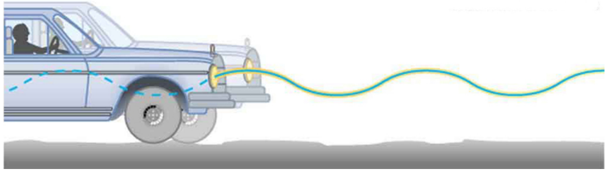
1. Enter the known values of *k* and *m*:
   Equation:
   $$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} = \frac{1}{2\pi}\sqrt{\frac{6.53 \times 10^4\ \text{N/m}}{900\ \text{kg}}}.$$
2. Calculate the frequency:
   Equation:
   $$\frac{1}{2\pi}\sqrt{72.6/\text{s}^{-2}} = 1.3656/\text{s}^{-1} \approx 1.36/\text{s}^{-1} = 1.36\ \text{Hz}.$$
3. You could use $T = 2\pi\sqrt{\frac{m}{k}}$ to calculate the period, but it is simpler to use the relationship $T = 1/f$ and substitute the value just found for f:
   Equation:
   $$T = \frac{1}{f} = \frac{1}{1.356\ \text{Hz}} = 0.738\ \text{s}.$$

Discussion

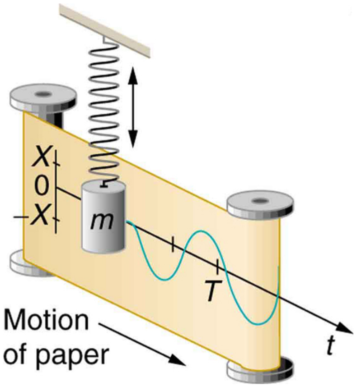
The values of T and f both seem about right for a bouncing car. You can observe these oscillations if you push down hard on the end of a car and let go.

The Link between Simple Harmonic Motion and Waves

If a time-exposure photograph of the bouncing car were taken as it drove by, the headlight would make a wavelike streak, as shown in [link]. Similarly, [link] shows an object bouncing on a spring as it leaves a wavelike "trace of its position on a moving strip of paper. Both waves are sine functions. All simple harmonic motion is intimately related to sine and cosine waves.



The bouncing car makes a wavelike motion. If the restoring force in the suspension system can be described only by Hooke’s law, then the wave is a sine function. (The wave is the trace produced by the headlight as the car moves to the right.)



The vertical position of an object bouncing on a spring is recorded on a strip of moving paper, leaving a sine wave.

The displacement as a function of time *t* in any simple harmonic motion—that is, one in which the net restoring force can be described by Hooke’s law, is given by

Equation:

$$x(t) = X \cos\frac{2\pi t}{T},$$

where X is amplitude. At $t=0$, the initial position is $x_0 = X$, and the displacement oscillates back and forth with a period T. (When $t=T$, we get $x=X$ again because cos $2\pi = 1$.). Furthermore, from this expression for x, the velocity v as a function of time is given by:

Equation:

$$v(t) = -v_{\text{max}} \sin\left(\frac{2\pi t}{T}\right),$$

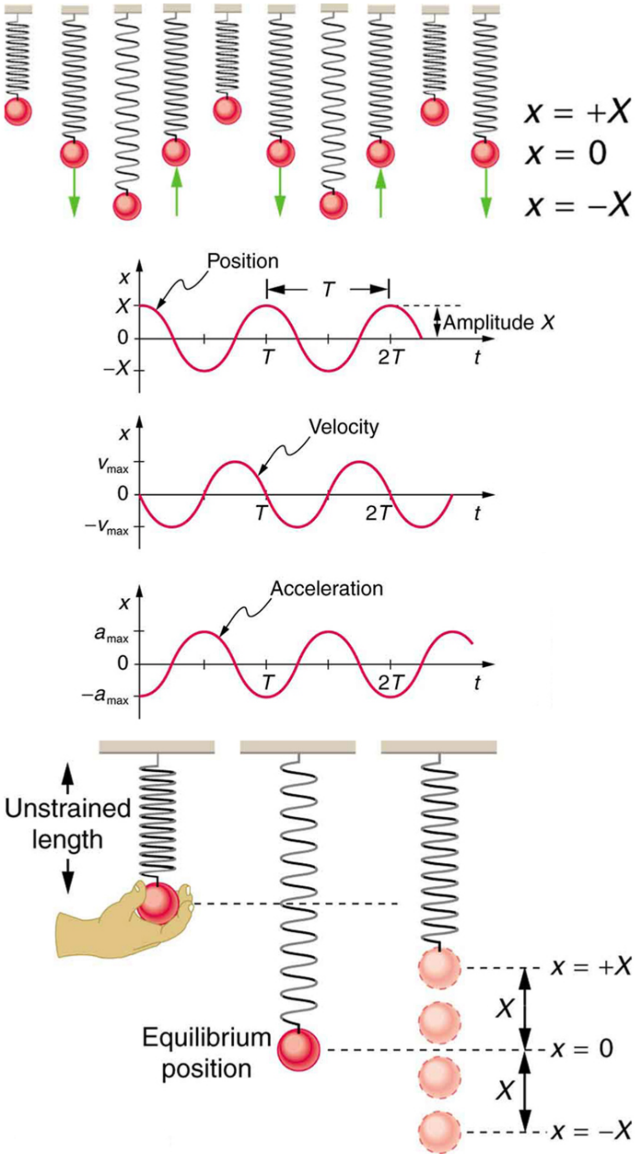
where $v_{\text{max}} = 2\text{π}X/T = X\sqrt{k/m}$. The object has zero velocity at maximum displacement—for example, $v = 0$ when $t = 0$, and at that time $x = X$. The minus sign in the first equation for $v(t)$ gives the correct direction for the velocity. Just after the start of the motion, for instance, the velocity is negative because the system is moving back toward the equilibrium point. Finally, we can get an expression for acceleration using Newton’s second law. [Then we have $x(t)$, $v(t)$, t, and $a(t)$, the quantities needed for kinematics and a description of simple harmonic motion.] According to Newton’s second law, the acceleration is $a = F/m = \text{kx}/m$. So, $a(t)$ is also a cosine function:

Equation:

$$a(t) = -\frac{\text{kX}}{m}\cos\frac{2\text{π}t}{T}.$$

Hence, $a(t)$ is directly proportional to and in the opposite direction to $x(t)$.

[link] shows the simple harmonic motion of an object on a spring and presents graphs of $x(t)$,$v(t\ )$, and $a(t)$ versus time.



Graphs of $x(t)$, $v(t)$, and $a(t)$ versus t for the motion of an object on a spring. The net force on the object can be described by Hooke's law, and so the object undergoes simple harmonic motion. Note that the initial position has the vertical displacement at its maximum value X; v is initially zero and then negative as the object moves down; and the initial acceleration is negative, back toward the equilibrium position and becomes zero at that point.

The most important point here is that these equations are mathematically straightforward and are valid for all simple harmonic motion. They are very useful in visualizing waves associated with simple harmonic motion, including visualizing how waves add with one another.

Exercise:
Check Your Understanding

Problem:

Suppose you pluck a banjo string. You hear a single note that starts out loud and slowly quiets over time. Describe what happens to the sound waves in terms of period, frequency and amplitude as the sound decreases in volume.

Solution:

Frequency and period remain essentially unchanged. Only amplitude decreases as volume decreases.

Exercise:
Check Your Understanding

Problem:

A babysitter is pushing a child on a swing. At the point where the swing reaches x, where would the corresponding point on a wave of this motion be located?

Solution:

x is the maximum deformation, which corresponds to the amplitude of the wave. The point on the wave would either be at the very top or the very bottom of the curve.

Note:
PhET Explorations: Masses and Springs
A realistic mass and spring laboratory. Hang masses from springs and adjust the spring stiffness and damping. You can even slow time. Transport the lab to different planets. A chart shows the kinetic, potential, and thermal energy for each spring.
https://phet.colorado.edu/sims/mass-spring-lab/mass-spring-lab_en.html

Section Summary

- Simple harmonic motion is oscillatory motion for a system that can be described only by Hooke’s law. Such a system is also called a simple harmonic oscillator.
- Maximum displacement is the amplitude X. The period T and frequency f of a simple harmonic oscillator are given by

  $T=2\pi\sqrt{\frac{m}{k}}$ and $f=\frac{1}{2\pi}\sqrt{\frac{k}{m}}$, where m is the mass of the system.
- Displacement in simple harmonic motion as a function of time is given by $x(t)=X\cos\frac{2\pi t}{T}$.
- The velocity is given by $v(t)=-v_{\max}\sin\frac{2\pi t}{T}$, where $v_{\max}=\sqrt{k/m}X$.
- The acceleration is found to be $a(t)=-\frac{kX}{m}\cos\frac{2\pi t}{T}$.

Conceptual Questions

Exercise:

Problem:

What conditions must be met to produce simple harmonic motion?

Exercise:

Problem:

(a) If frequency is not constant for some oscillation, can the oscillation be simple harmonic motion?

(b) Can you think of any examples of harmonic motion where the frequency may depend on the amplitude?

Exercise:

Problem:

Give an example of a simple harmonic oscillator, specifically noting how its frequency is independent of amplitude.

Exercise:

Problem:

Explain why you expect an object made of a stiff material to vibrate at a higher frequency than a similar object made of a spongy material.

Exercise:

Problem:

As you pass a freight truck with a trailer on a highway, you notice that its trailer is bouncing up and down slowly. Is it more likely that the trailer is heavily loaded or nearly empty? Explain your answer.

Exercise:

Problem:

Some people modify cars to be much closer to the ground than when manufactured. Should they install stiffer springs? Explain your answer.

Problems & Exercises

Exercise:

Problem:

A type of cuckoo clock keeps time by having a mass bouncing on a spring, usually something cute like a cherub in a chair. What force constant is needed to produce a period of 0.500 s for a 0.0150-kg mass?

Solution:

2.37 N/m

Exercise:

Problem:

If the spring constant of a simple harmonic oscillator is doubled, by what factor will the mass of the system need to change in order for the frequency of the motion to remain the same?

Exercise:

Problem:

A 0.500-kg mass suspended from a spring oscillates with a period of 1.50 s. How much mass must be added to the object to change the period to 2.00 s?

Solution:

0.389 kg

Exercise:

Problem:

By how much leeway (both percentage and mass) would you have in the selection of the mass of the object in the previous problem if you did not wish the new period to be greater than 2.01 s or less than 1.99 s?

Exercise:

Problem:

Suppose you attach the object with mass m to a vertical spring originally at rest, and let it bounce up and down. You release the object from rest at the spring’s original rest length. (a) Show that the spring exerts an upward force of $2.00\ \text{mg}$ on the object at its lowest point. (b) If the spring has a force constant of $10.0\ \text{N}/\text{m}$ and a 0.25-kg-mass object is set in motion as described, find the amplitude of the oscillations. (c) Find the maximum velocity.

Exercise:

Problem:

A diver on a diving board is undergoing simple harmonic motion. Her mass is 55.0 kg and the period of her motion is 0.800 s. The next diver is a male whose period of simple harmonic oscillation is 1.05 s. What is his mass if the mass of the board is negligible?

Solution:

94.7 kg

Exercise:

Problem:

Suppose a diving board with no one on it bounces up and down in a simple harmonic motion with a frequency of 4.00 Hz. The board has an effective mass of 10.0 kg. What is the frequency of the simple harmonic motion of a 75.0-kg diver on the board?

Exercise:

Problem:

This child’s toy relies on springs to keep infants entertained. (credit: By Humboldthead, Flickr)

The device pictured in [link] entertains infants while keeping them from wandering. The child bounces in a harness suspended from a door frame by a spring constant.

(a) If the spring stretches 0.250 m while supporting an 8.0-kg child, what is its spring constant?

(b) What is the time for one complete bounce of this child? (c) What is the child’s maximum velocity if the amplitude of her bounce is 0.200 m?

Exercise:

Problem:

A 90.0-kg skydiver hanging from a parachute bounces up and down with a period of 1.50 s. What is the new period of oscillation when a second skydiver, whose mass is 60.0 kg, hangs from the legs of the first, as seen in [link].

The oscillations of one skydiver are about to be affected by a second skydiver. (credit: U.S. Army, www.army.mil)

Solution:

1.94 s

Glossary

amplitude
: the maximum displacement from the equilibrium position of an object oscillating around the equilibrium position

simple harmonic motion

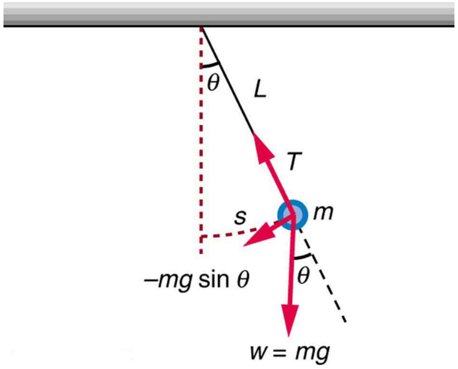
the oscillatory motion in a system where the net force can be described by Hooke’s law

simple harmonic oscillator

a device that implements Hooke’s law, such as a mass that is attached to a spring, with the other end of the spring being connected to a rigid support such as a wall

The Simple Pendulum

- Measure acceleration due to gravity.



A simple pendulum has a small-diameter bob and a string that has a very small mass but is strong enough not to stretch appreciably. The linear displacement from equilibrium is s, the length of the arc. Also shown are the forces on the bob, which result in a net force of $-\text{mg}\ \text{sin}\theta$ toward the equilibrium position—that is, a restoring force.

Pendulums are in common usage. Some have crucial uses, such as in clocks; some are for fun, such as a child’s swing; and some are just there, such as the sinker on a fishing line. For small displacements, a pendulum is a simple harmonic oscillator. A **simple pendulum** is defined to have an object that has a small mass, also known as the pendulum bob, which is suspended from a light wire or string, such as shown in [link]. Exploring the simple pendulum a bit further, we can discover the conditions under which it performs simple harmonic motion, and we can derive an interesting expression for its period.

We begin by defining the displacement to be the arc length s. We see from [link] that the net force on the bob is tangent to the arc and equals $-\text{mg}\ \text{sin}\ \theta$. (The weight mg has components $\text{mg}\ \text{cos}\ \theta$ along the string and $\text{mg}\ \text{sin}\ \theta$ tangent to the arc.) Tension in the string exactly cancels the component $\text{mg}\ \text{cos}\ \theta$ parallel to the string. This leaves a *net* restoring force back toward the equilibrium position at $\theta = 0$.

Now, if we can show that the restoring force is directly proportional to the displacement, then we have a simple harmonic oscillator. In trying to determine if we have a simple harmonic oscillator, we should note that for small angles (less than about 15^{o}), $\text{sin}\ \theta \approx \ \theta$ ($\text{sin}\ \theta$ and θ differ by about 1% or less at smaller angles). Thus, for angles less than about 15^{o}, the restoring force F is

Equation:

$$F \approx -\text{mg}\theta.$$

The displacement s is directly proportional to θ. When θ is expressed in radians, the arc length in a circle is related to its radius (L in this instance) by:

Equation:

$$s = L\theta,$$

so that

Equation:

$$\theta = \frac{s}{L}.$$

For small angles, then, the expression for the restoring force is:
Equation:

$$F\approx -\frac{\text{mg}}{L}s$$

This expression is of the form:
Equation:

$$F=-\text{kx},$$

where the force constant is given by $k=mg/L$ and the displacement is given by $x=s$. For angles less than about 15º, the restoring force is directly proportional to the displacement, and the simple pendulum is a simple harmonic oscillator.

Using this equation, we can find the period of a pendulum for amplitudes less than about 15º. For the simple pendulum:
Equation:

$$T=2\pi\sqrt{\frac{m}{k}}=2\pi\sqrt{\frac{m}{\text{mg}/L}}.$$

Thus,
Equation:

$$T=2\pi\sqrt{\frac{L}{g}}$$

for the period of a simple pendulum. This result is interesting because of its simplicity. The only things that affect the period of a simple pendulum are its length and the acceleration due to gravity. The period is completely independent of other factors, such as mass. As with simple harmonic oscillators, the period T for a pendulum is nearly independent of amplitude, especially if θ is less than about 15º. Even simple pendulum clocks can be finely adjusted and accurate.

Note the dependence of T on g. If the length of a pendulum is precisely known, it can actually be used to measure the acceleration due to gravity. Consider the following example.

Example:
Measuring Acceleration due to Gravity: The Period of a Pendulum
What is the acceleration due to gravity in a region where a simple pendulum having a length 75.000 cm has a period of 1.7357 s?
Strategy
We are asked to find g given the period T and the length L of a pendulum. We can solve $T = 2\pi\sqrt{\frac{L}{g}}$ for g, assuming only that the angle of deflection is less than 15º.
Solution

1. Square $T = 2\pi\sqrt{\frac{L}{g}}$ and solve for g:
   Equation:
   $$g = 4\pi^2 \frac{L}{T^2}.$$
2. Substitute known values into the new equation:
   Equation:
   $$g = 4\pi^2 \frac{0.75000\text{ m}}{(1.7357\text{ s})^2}.$$
3. Calculate to find g:
   Equation:
   $$g = 9.8281\text{ m/s}^2.$$

Discussion

This method for determining *g* can be very accurate. This is why length and period are given to five digits in this example. For the precision of the approximation $\sin\theta \approx \theta$ to be better than the precision of the pendulum length and period, the maximum displacement angle should be kept below about 0.5°.

Note:

Making Career Connections

Knowing *g* can be important in geological exploration; for example, a map of *g* over large geographical regions aids the study of plate tectonics and helps in the search for oil fields and large mineral deposits.

Note:

Take Home Experiment: Determining *g*

Use a simple pendulum to determine the acceleration due to gravity *g* in your own locale. Cut a piece of a string or dental floss so that it is about 1 m long. Attach a small object of high density to the end of the string (for example, a metal nut or a car key). Starting at an angle of less than 10°, allow the pendulum to swing and measure the pendulum’s period for 10 oscillations using a stopwatch. Calculate *g*. How accurate is this measurement? How might it be improved?

Exercise:

Check Your Understanding

Problem:

An engineer builds two simple pendula. Both are suspended from small wires secured to the ceiling of a room. Each pendulum hovers 2 cm above the floor. Pendulum 1 has a bob with a mass of 10 kg. Pendulum 2 has a bob with a mass of 100 kg. Describe how the motion of the pendula will differ if the bobs are both displaced by 12^{o}.

Solution:

The movement of the pendula will not differ at all because the mass of the bob has no effect on the motion of a simple pendulum. The pendula are only affected by the period (which is related to the pendulum's length) and by the acceleration due to gravity.

Note:
PhET Explorations: Pendulum Lab
Play with one or two pendulums and discover how the period of a simple pendulum depends on the length of the string, the mass of the pendulum bob, and the amplitude of the swing. It's easy to measure the period using the photogate timer. You can vary friction and the strength of gravity. Use the pendulum to find the value of g on planet X. Notice the anharmonic behavior at large amplitude.
https://phet.colorado.edu/sims/pendulum-lab/pendulum-lab_en.html

Section Summary

- A mass m suspended by a wire of length L is a simple pendulum and undergoes simple harmonic motion for amplitudes less than about 15^{o}.

  The period of a simple pendulum is
  Equation:

$$T = 2\pi\sqrt{\frac{L}{g}},$$

where L is the length of the string and g is the acceleration due to gravity.

Conceptual Questions

Exercise:

Problem:

Pendulum clocks are made to run at the correct rate by adjusting the pendulum’s length. Suppose you move from one city to another where the acceleration due to gravity is slightly greater, taking your pendulum clock with you, will you have to lengthen or shorten the pendulum to keep the correct time, other factors remaining constant? Explain your answer.

Problems & Exercises

As usual, the acceleration due to gravity in these problems is taken to be $g = 9.80 \text{ m/s}^2$, unless otherwise specified.

Exercise:

Problem:

What is the length of a pendulum that has a period of 0.500 s?

Solution:

6.21 cm

Exercise:

Problem:

Some people think a pendulum with a period of 1.00 s can be driven with “mental energy” or psycho kinetically, because its period is the same as an average heartbeat. True or not, what is the length of such a pendulum?

Exercise:

Problem: What is the period of a 1.00-m-long pendulum?

Solution:

2.01 s

Exercise:

Problem:

How long does it take a child on a swing to complete one swing if her center of gravity is 4.00 m below the pivot?

Exercise:

Problem:

The pendulum on a cuckoo clock is 5.00 cm long. What is its frequency?

Solution:

2.23 Hz

Exercise:

Problem:

Two parakeets sit on a swing with their combined center of mass 10.0 cm below the pivot. At what frequency do they swing?

Exercise:

Problem:

(a) A pendulum that has a period of 3.00000 s and that is located where the acceleration due to gravity is $9.79\ \mathrm{m/s}^2$ is moved to a location where it the acceleration due to gravity is $9.82\ \mathrm{m/s}^2$. What is its new period? (b) Explain why so many digits are needed in the value for the period, based on the relation between the period and the acceleration due to gravity.

Solution:

(a) 2.99541 s

(b) Since the period is related to the square root of the acceleration of gravity, when the acceleration changes by 1% the period changes by $(0.01)^2 = 0.01\%$ so it is necessary to have at least 4 digits after the decimal to see the changes.

Exercise:

Problem:

A pendulum with a period of 2.00000 s in one location $\left(g = 9.80\ \mathrm{m/s}^2\right)$ is moved to a new location where the period is now 1.99796 s. What is the acceleration due to gravity at its new location?

Exercise:

Problem:

(a) What is the effect on the period of a pendulum if you double its length?

(b) What is the effect on the period of a pendulum if you decrease its length by 5.00%?

Solution:

(a) Period increases by a factor of 1.41 ($\sqrt{2}$)

(b) Period decreases to 97.5% of old period

Exercise:

Problem:

Find the ratio of the new/old periods of a pendulum if the pendulum were transported from Earth to the Moon, where the acceleration due to gravity is 1.63 m/s^2.

Exercise:

Problem:

At what rate will a pendulum clock run on the Moon, where the acceleration due to gravity is 1.63 m/s^2, if it keeps time accurately on Earth? That is, find the time (in hours) it takes the clock's hour hand to make one revolution on the Moon.

Solution:

Slow by a factor of 2.45

Exercise:

Problem:

Suppose the length of a clock's pendulum is changed by 1.000%, exactly at noon one day. What time will it read 24.00 hours later, assuming it the pendulum has kept perfect time before the change? Note that there are two answers, and perform the calculation to four-digit precision.

Exercise:

Problem:

If a pendulum-driven clock gains 5.00 s/day, what fractional change in pendulum length must be made for it to keep perfect time?

Solution:

length must increase by 0.0116%.

Glossary

simple pendulum

an object with a small mass suspended from a light wire or string

Energy and the Simple Harmonic Oscillator

- Determine the maximum speed of an oscillating system.

To study the energy of a simple harmonic oscillator, we first consider all the forms of energy it can have We know from Hooke’s Law: Stress and Strain Revisited that the energy stored in the deformation of a simple harmonic oscillator is a form of potential energy given by:

Equation:

$$\mathrm{PE_{el}} = \frac{1}{2}kx^2.$$

Because a simple harmonic oscillator has no dissipative forces, the other important form of energy is kinetic energy KE. Conservation of energy for these two forms is:

Equation:

$$\mathrm{KE} + \mathrm{PE_{el}} = \mathrm{constant}$$

or

Equation:

$$\frac{1}{2}\mathrm{mv}^2 + \frac{1}{2}\mathrm{kx}^2 = \mathrm{constant.}$$

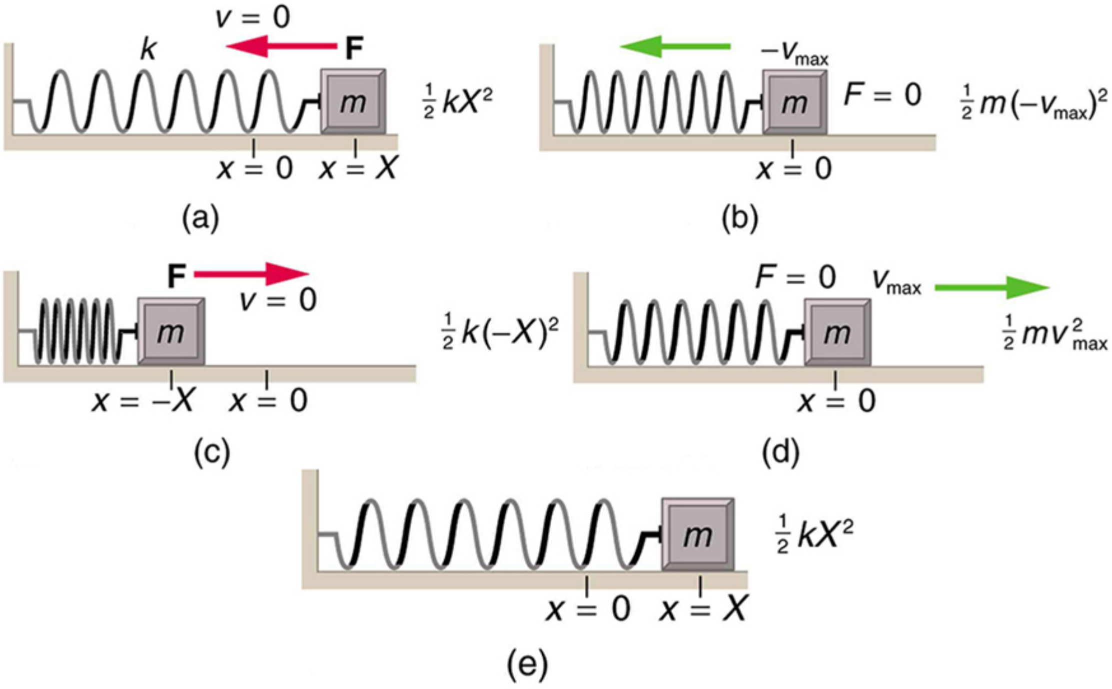
This statement of conservation of energy is valid for *all* simple harmonic oscillators, including ones where the gravitational force plays a role

Namely, for a simple pendulum we replace the velocity with $v = L\omega$, the spring constant with $k = \mathrm{mg}/L$, and the displacement term with $x = L\theta$. Thus

Equation:

$$\frac{1}{2}\mathrm{mL}^2\omega^2 + \frac{1}{2}\mathrm{mgL}\theta^2 = \mathrm{constant.}$$

In the case of undamped simple harmonic motion, the energy oscillates back and forth between kinetic and potential, going completely from one to the other as the system oscillates. So for the simple example of an object on a frictionless surface attached to a spring, as shown again in [link], the motion starts with all of the energy stored in the spring. As the object starts to move, the elastic potential energy is converted to kinetic energy, becoming entirely kinetic energy at the equilibrium position. It is then converted back into elastic potential energy by the spring, the velocity becomes zero when the kinetic energy is completely converted, and so on. This concept provides extra insight here and in later applications of simple harmonic motion, such as alternating current circuits.



The transformation of energy in simple harmonic motion is illustrated for an object attached to a spring on a frictionless surface.

The conservation of energy principle can be used to derive an expression for velocity v. If we start our simple harmonic motion with zero velocity and maximum displacement ($x = X$), then the total energy is

Equation:

$$\frac{1}{2}\mathrm{kX}^2.$$

This total energy is constant and is shifted back and forth between kinetic energy and potential energy, at most times being shared by each. The conservation of energy for this system in equation form is thus:

Equation:

$$\frac{1}{2}\mathrm{mv}^2+\frac{1}{2}\mathrm{kx}^2=\frac{1}{2}\mathrm{kX}^2.$$

Solving this equation for v yields:

Equation:

$$v=\pm\sqrt{\frac{k}{m}(X^2-x^2)}.$$

Manipulating this expression algebraically gives:

Equation:

$$v=\pm\sqrt{\frac{k}{m}}X\sqrt{1-\frac{x^2}{X^2}}$$

and so

Equation:

$$v=\pm v_{\mathrm{max}}\sqrt{1-\frac{x^2}{X^2}},$$

where

Equation:

$$v_{\max} = \sqrt{\frac{k}{m}}X.$$

From this expression, we see that the velocity is a maximum ($v_{\max}$) at $x = 0$, as stated earlier in $v(t) = -v_{\max}\ \sin\ \frac{2\pi t}{T}$. Notice that the maximum velocity depends on three factors. Maximum velocity is directly proportional to amplitude. As you might guess, the greater the maximum displacement the greater the maximum velocity. Maximum velocity is also greater for stiffer systems, because they exert greater force for the same displacement. This observation is seen in the expression for $v_{\max}$; it is proportional to the square root of the force constant k. Finally, the maximum velocity is smaller for objects that have larger masses, because the maximum velocity is inversely proportional to the square root of m. For a given force, objects that have large masses accelerate more slowly.

A similar calculation for the simple pendulum produces a similar result, namely:

Equation:

$$\omega_{\max} = \sqrt{\frac{g}{L}}\theta_{\max}.$$

Example:

Determine the Maximum Speed of an Oscillating System: A Bumpy Road

Suppose that a car is 900 kg and has a suspension system that has a force constant $k = 6.53 \times 10^4\ \text{N/m}$. The car hits a bump and bounces with an amplitude of 0.100 m. What is its maximum vertical velocity if you assume no damping occurs?

Strategy

We can use the expression for $v_{\max}$ given in $v_{\max} = \sqrt{\frac{k}{m}}X$ to determine the maximum vertical velocity. The variables m and k are given in the problem statement, and the maximum displacement X is 0.100 m.

Solution

1. Identify known.
2. Substitute known values into $v_{\text{max}} = \sqrt{\frac{k}{m}}X$:

   Equation:

   $$v_{\text{max}} = \sqrt{\frac{6.53 \times 10^4 \text{ N/m}}{900 \text{ kg}}}(0.100 \text{ m}).$$

3. Calculate to find v_{max}= 0.852 m/s.

Discussion

This answer seems reasonable for a bouncing car. There are other ways to use conservation of energy to find v_{max}. We could use it directly, as was done in the example featured in Hooke's Law: Stress and Strain Revisited.

The small vertical displacement y of an oscillating simple pendulum, starting from its equilibrium position, is given as

Equation:

$$y(t) = a \sin \omega t,$$

where a is the amplitude, ω is the angular velocity and t is the time taken. Substituting $\omega = \frac{2\pi}{T}$, we have

Equation:

$$y(t) = a \sin\left(\frac{2\pi t}{T}\right).$$

Thus, the displacement of pendulum is a function of time as shown above.

Also the velocity of the pendulum is given by

Equation:

$$v(t) = \frac{2a\pi}{T} \cos\left(\frac{2\pi t}{T}\right),$$

so the motion of the pendulum is a function of time.

Exercise:
Check Your Understanding

Problem:

Why does it hurt more if your hand is snapped with a ruler than with a loose spring, even if the displacement of each system is equal?

Solution:

The ruler is a stiffer system, which carries greater force for the same amount of displacement. The ruler snaps your hand with greater force, which hurts more.

Exercise:
Check Your Understanding

Problem:

You are observing a simple harmonic oscillator. Identify one way you could decrease the maximum velocity of the system.

Solution:

You could increase the mass of the object that is oscillating.

Section Summary

- Energy in the simple harmonic oscillator is shared between elastic potential energy and kinetic energy, with the total being constant:
  Equation:

$$\frac{1}{2}\text{mv}^2+\frac{1}{2}\text{kx}^2=\text{constant}.$$

- Maximum velocity depends on three factors: it is directly proportional to amplitude, it is greater for stiffer systems, and it is smaller for objects that have larger masses:
  Equation:

$$v_{\text{max}} = \sqrt{\frac{k}{m}}X.$$

Conceptual Questions

Exercise:

Problem:

Explain in terms of energy how dissipative forces such as friction reduce the amplitude of a harmonic oscillator. Also explain how a driving mechanism can compensate. (A pendulum clock is such a system.)

Problems & Exercises

Exercise:

Problem:

The length of nylon rope from which a mountain climber is suspended has a force constant of $1.40 \times 10^4 \text{ N/m}$.

(a) What is the frequency at which he bounces, given his mass plus and the mass of his equipment are 90.0 kg?

(b) How much would this rope stretch to break the climber’s fall if he free-falls 2.00 m before the rope runs out of slack? Hint: Use conservation of energy.

(c) Repeat both parts of this problem in the situation where twice this length of nylon rope is used.

Solution:

(a) 1.99 Hz

(b) 50.2 cm

(c) 1.41 Hz, 0.710 m

Exercise:

Problem: **Engineering Application**

Near the top of the Citigroup Center building in New York City, there is an object with mass of 4.00×10^{5} kg on springs that have adjustable force constants. Its function is to dampen wind-driven oscillations of the building by oscillating at the same frequency as the building is being driven—the driving force is transferred to the object, which oscillates instead of the entire building. (a) What effective force constant should the springs have to make the object oscillate with a period of 2.00 s? (b) What energy is stored in the springs for a 2.00-m displacement from equilibrium?

Solution:

(a) $3.95\times 10^{6}\text{ N/m}$

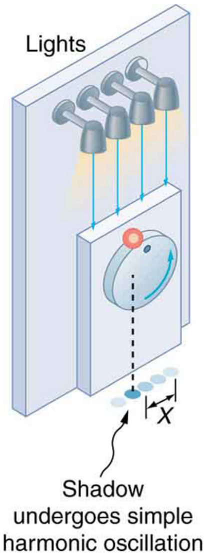
(b) $7.90\times 10^{6}\text{ J}$

Uniform Circular Motion and Simple Harmonic Motion

- Compare simple harmonic motion with uniform circular motion.

The horses on this merry-go-round exhibit uniform circular motion. (credit: Wonderlane, Flickr)

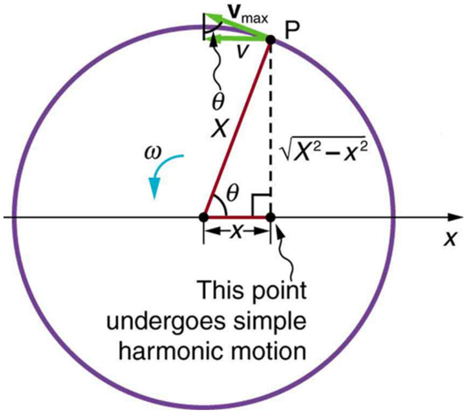
There is an easy way to produce simple harmonic motion by using uniform circular motion. [link] shows one way of using this method. A ball is attached to a uniformly rotating vertical turntable, and its shadow is projected on the floor as shown. The shadow undergoes simple harmonic motion. Hooke’s law usually describes uniform circular motions (ω constant) rather than systems that have large visible displacements. So observing the projection of uniform circular motion, as in [link], is often easier than observing a precise large-scale simple harmonic oscillator. If studied in sufficient depth, simple harmonic motion produced in this manner can give considerable insight into many aspects of oscillations and waves and is very useful mathematically. In our brief treatment, we shall indicate some of the major features of this relationship and how they might be useful.



The shadow of a ball rotating at constant angular velocity ω on a turntable goes back and forth in precise simple harmonic motion.

[link] shows the basic relationship between uniform circular motion and simple harmonic motion. The point P travels around the circle at constant angular velocity ω. The point P is analogous to an object on the merry-go-

round. The projection of the position of P onto a fixed axis undergoes simple harmonic motion and is analogous to the shadow of the object. At the time shown in the figure, the projection has position x and moves to the left with velocity v. The velocity of the point P around the circle equals v_{max}.The projection of v_{max} on the x-axis is the velocity v of the simple harmonic motion along the x-axis.


$\mathbf{v}_{\text{max}}$
P
v
θ
X
ω
$\sqrt{X^2-x^2}$
θ
x
x
This point undergoes simple harmonic motion

A point P moving on a circular path with a constant angular velocity ω is undergoing uniform circular motion. Its projection on the x-axis undergoes simple harmonic motion. Also shown is the velocity of this point around the circle, v_{max}, and its projection, which is v. Note that these velocities form a similar triangle to the displacement triangle.

To see that the projection undergoes simple harmonic motion, note that its position x is given by

Equation:

$$x = X\cos\theta,$$

where $\theta = \omega t$, ω is the constant angular velocity, and X is the radius of the circular path. Thus,

Equation:

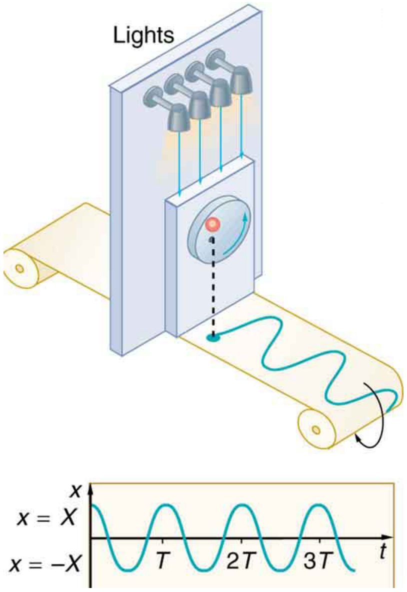
$$x = X\cos\omega t.$$

The angular velocity ω is in radians per unit time; in this case 2π radians is the time for one revolution T. That is, $\omega = 2\pi/T$. Substituting this expression for ω, we see that the position x is given by:

Equation:

$$x(t) = \cos\left(\frac{2\pi t}{T}\right).$$

This expression is the same one we had for the position of a simple harmonic oscillator in Simple Harmonic Motion: A Special Periodic Motion. If we make a graph of position versus time as in [link], we see again the wavelike character (typical of simple harmonic motion) of the projection of uniform circular motion onto the x-axis.



The position of the projection of uniform circular motion performs simple harmonic motion, as this wavelike graph of x versus t indicates.

Now let us use [link] to do some further analysis of uniform circular motion as it relates to simple harmonic motion. The triangle formed by the velocities in the figure and the triangle formed by the displacements (X, x, and $\sqrt{X^2 - x^2}$) are similar right triangles. Taking ratios of similar sides, we see that

Equation:

$$\frac{v}{v_{\text{max}}} = \frac{\sqrt{X^2 - x^2}}{X} = \sqrt{1 - \frac{x^2}{X^2}}.$$

We can solve this equation for the speed v or
Equation:

$$v = v_{\text{max}}\sqrt{1 - \frac{x^2}{X^2}}.$$

This expression for the speed of a simple harmonic oscillator is exactly the same as the equation obtained from conservation of energy considerations in Energy and the Simple Harmonic Oscillator.You can begin to see that it is possible to get all of the characteristics of simple harmonic motion from an analysis of the projection of uniform circular motion.

Finally, let us consider the period T of the motion of the projection. This period is the time it takes the point P to complete one revolution. That time is the circumference of the circle $2\pi X$ divided by the velocity around the circle, v_{max}. Thus, the period T is
Equation:

$$T = \frac{2\pi\text{X}}{v_{\text{max}}}.$$

We know from conservation of energy considerations that
Equation:

$$v_{\text{max}} = \sqrt{\frac{k}{m}}X.$$

Solving this equation for X/v_{max} gives
Equation:

$$\frac{X}{v_{\text{max}}} = \sqrt{\frac{m}{k}}.$$

Substituting this expression into the equation for T yields

Equation:

$$T = 2\pi\sqrt{\frac{m}{k}}.$$

Thus, the period of the motion is the same as for a simple harmonic oscillator. We have determined the period for any simple harmonic oscillator using the relationship between uniform circular motion and simple harmonic motion.

Some modules occasionally refer to the connection between uniform circular motion and simple harmonic motion. Moreover, if you carry your study of physics and its applications to greater depths, you will find this relationship useful. It can, for example, help to analyze how waves add when they are superimposed.

Exercise:

Check Your Understanding

Problem:

Identify an object that undergoes uniform circular motion. Describe how you could trace the simple harmonic motion of this object as a wave.

Solution:

A record player undergoes uniform circular motion. You could attach dowel rod to one point on the outside edge of the turntable and attach a pen to the other end of the dowel. As the record player turns, the pen will move. You can drag a long piece of paper under the pen, capturing its motion as a wave.

Section Summary

A projection of uniform circular motion undergoes simple harmonic oscillation.

Problems & Exercises

Exercise:

Problem:

(a)What is the maximum velocity of an 85.0-kg person bouncing on a bathroom scale having a force constant of $1.50 \times 10^6 \text{ N/m}$, if the amplitude of the bounce is 0.200 cm? (b)What is the maximum energy stored in the spring?

Solution:

a). 0.266 m/s

b). 3.00 J

Exercise:

Problem:

A novelty clock has a 0.0100-kg mass object bouncing on a spring that has a force constant of 1.25 N/m. What is the maximum velocity of the object if the object bounces 3.00 cm above and below its equilibrium position? (b) How many joules of kinetic energy does the object have at its maximum velocity?

Exercise:

Problem:

At what positions is the speed of a simple harmonic oscillator half its maximum? That is, what values of x/X give $v = \pm v_{\text{max}}/2$, where X is the amplitude of the motion?

Solution:

$\pm\frac{\sqrt{3}}{2}$

Exercise:

Problem:

A ladybug sits 12.0 cm from the center of a Beatles music album spinning at 33.33 rpm. What is the maximum velocity of its shadow on the wall behind the turntable, if illuminated parallel to the record by the parallel rays of the setting Sun?

Damped Harmonic Motion

- Compare and discuss underdamped and overdamped oscillating systems.
- Explain critically damped system.

In order to counteract dampening forces, this mom needs to keep pushing the swing. (credit: Mohd Fazlin Mohd Effendy Ooi, Flickr)

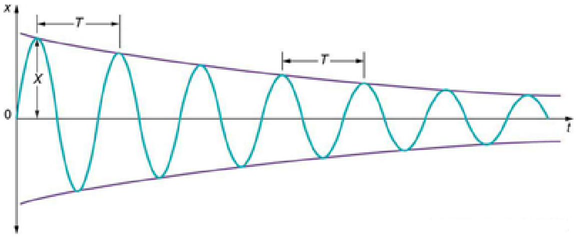
A guitar string stops oscillating a few seconds after being plucked. To keep a child happy on a swing, you must keep pushing. Although we can often make friction and other non-conservative forces negligibly small, completely undamped motion is rare. In fact, we may even want to damp oscillations, such as with car shock absorbers.

For a system that has a small amount of damping, the period and frequency are nearly the same as for simple harmonic motion, but the amplitude gradually decreases as shown in [link]. This occurs because the non-conservative damping force removes energy from the system, usually in the form of thermal energy. In general, energy removal by non-conservative forces is described as

Equation:

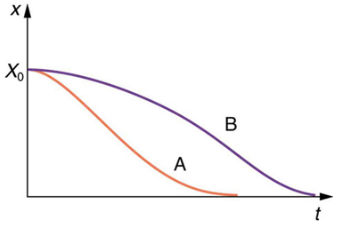
$$W_{\text{nc}} = \Delta(\text{KE} + \text{PE}),$$

where W_{nc} is work done by a non-conservative force (here the damping force). For a damped harmonic oscillator, W_{nc} is negative because it removes mechanical energy (KE + PE) from the system.



In this graph of displacement versus time for a harmonic oscillator with a small amount of damping, the amplitude slowly decreases, but the period and frequency are nearly the same as if the system were completely undamped.

If you gradually *increase* the amount of damping in a system, the period and frequency begin to be affected, because damping opposes and hence slows the back and forth motion. (The net force is smaller in both directions.) If there is very large damping, the system does not even oscillate—it slowly moves toward equilibrium. [link] shows the displacement of a harmonic oscillator for different amounts of damping. When we want to damp out oscillations, such as in the suspension of a car, we may want the system to return to equilibrium as quickly as possible **Critical damping** is defined as the condition in which the damping of an oscillator results in it returning as quickly as possible to its equilibrium position The critically damped system may overshoot the equilibrium position, but if it does, it will do so only once. Critical damping is represented by Curve A in [link]. With less-than critical damping, the system will return to equilibrium faster but will overshoot and cross over one or more times. Such a system is **underdamped**; its displacement is represented by the curve in [link]. Curve B in [link] represents an **overdamped** system. As with critical damping, it too may overshoot the equilibrium position, but will reach equilibrium over a longer period of time.



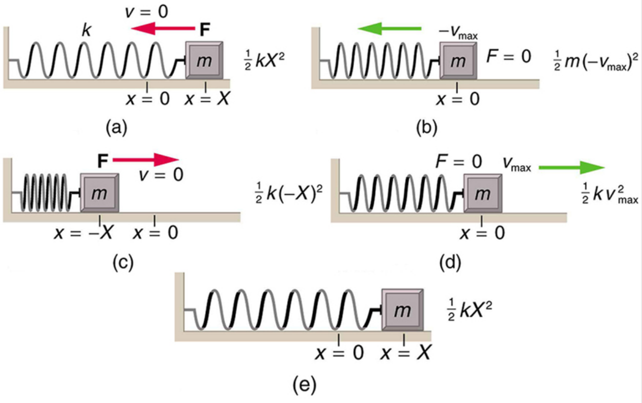
Displacement versus time for a critically damped harmonic oscillator (A) and an overdamped harmonic oscillator (B). The critically damped oscillator returns to equilibrium at $X = 0$ in the smallest time possible without overshooting.

Critical damping is often desired, because such a system returns to equilibrium rapidly and remains at equilibrium as well. In addition, a constant force applied to a critically damped system moves the system to a new equilibrium position in the shortest time possible without overshooting or oscillating about the new position. For example, when you stand on bathroom scales that have a needle gauge, the needle moves to its equilibrium position without oscillating. It would be quite inconvenient if the needle oscillated about the new equilibrium position for a long time before settling. Damping forces can vary greatly in character. Friction, for example, is sometimes independent of velocity (as assumed in most places in this text). But many damping forces depend on velocity—sometimes in complex ways, sometimes simply being proportional to velocity.

Example:

Damping an Oscillatory Motion: Friction on an Object Connected to a Spring

Damping oscillatory motion is important in many systems, and the ability to control the damping is even more so. This is generally attained using non-conservative forces such as the friction between surfaces, and viscosity for objects moving through fluids. The following example considers friction. Suppose a 0.200-kg object is connected to a spring as shown in [link], but there is simple friction between the object and the surface, and the coefficient of friction μ_k is equal to 0.0800. (a) What is the frictional force between the surfaces? (b) What total distance does the object travel if it is released 0.100 m from equilibrium, starting at $v = 0$? The force constant of the spring is $k = 50.0\ \text{N/m}$.



The transformation of energy in simple harmonic motion is illustrated for an object attached to a spring on a frictionless surface.

Strategy

This problem requires you to integrate your knowledge of various concepts regarding waves, oscillations, and damping. To solve an integrated concept problem, you must first identify the physical principles involved. Part (a) is about the frictional force. This is a topic involving the application of Newton's Laws. Part (b) requires an understanding of work and conservation of energy, as well as some understanding of horizontal oscillatory systems.

Now that we have identified the principles we must apply in order to solve the problems, we need to identify the knowns and unknowns for each part of the question, as well as the quantity that is constant in Part (a) and Part (b) of the question.

Solution a

1. Choose the proper equation: Friction is $f = \mu_{\text{k}}\text{mg}$.
2. Identify the known values.
3. Enter the known values into the equation:
   Equation:

$$f = (0.0800)(0.200\ \text{kg})(9.80\ \text{m/s}^2).$$

4. Calculate and convert units: $f = 0.157\ \text{N}$.

Discussion a

The force here is small because the system and the coefficients are small.

Solution b

Identify the known:

- The system involves elastic potential energy as the spring compresses and expands, friction that is related to the work done, and the kinetic energy as the body speeds up and slows down.
- Energy is not conserved as the mass oscillates because friction is a non-conservative force.
- The motion is horizontal, so gravitational potential energy does not need to be considered.
- Because the motion starts from rest, the energy in the system is initially $\mathrm{PE_{el,i}} = (1/2)\mathrm{kX}^2$. This energy is removed by work done by friction $W_{\mathrm{nc}} = -\mathrm{fd}$, where d is the total distance traveled and $f = \mu_{\mathrm{k}}\mathrm{mg}$ is the force of friction. When the system stops moving, the friction force will balance the force exerted by the spring, so $\mathrm{PE_{e1,f}} = (1/2)kx^2$ where x is the final position and is given by
  Equation:

$$\begin{aligned} F_{\mathrm{el}} &= f \\ \mathrm{kx} &= \mu_{\mathrm{k}}\mathrm{mg}. \\ x &= \frac{\mu_{\mathrm{k}}\mathrm{mg}}{k} \end{aligned}$$

1. By equating the work done to the energy removed, solve for the distance d .
2. The work done by the non-conservative forces equals the initial, stored elastic potential energy. Identify the correct equation to use:
   Equation:

$$\mathrm{W_{nc}} = \mathit{\Delta}(\mathrm{KE}+\mathrm{PE}) = \mathrm{PE_{el,f}} - \mathrm{PE_{el,i}} = \frac{1}{2}k\left(\left(\frac{\mu_{\mathrm{k}}mg}{k}\right)^2 - X^2\right).$$

3. Recall that $W_{\mathrm{nc}} = -\mathrm{fd}$.
4. Enter the friction as $f = \mu_{\mathrm{k}}\mathrm{mg}$ into $W_{\mathrm{nc}} = -\mathrm{fd}$, thus
   Equation:

$$W_{\mathrm{nc}} = -\mu_{\mathrm{k}}\mathrm{mgd}.$$

5. Combine these two equations to find
   Equation:

$$\frac{1}{2}k\left(\left(\frac{\mu_{\mathrm{k}}mg}{k}\right)^2 - X^2\right) = -\mu_{\mathrm{k}}mgd.$$

6. Solve the equation for d :
   Equation:

$$d = \frac{\mathrm{k}}{2\mu_{\mathrm{k}}\mathrm{mg}}\left(X^2 - \left(\frac{\mu_{\mathrm{k}}\mathrm{mg}}{k}\right)^2\right).$$

7. Enter the known values into the resulting equation:
   Equation:

$$d = \frac{50.0\ \mathrm{N/m}}{2(0.0800)(0.200\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)} \quad (0.100\ \mathrm{m})^2 - \quad {\frac{(0.0800)(0.200\ \mathrm{kg})\left(9.80\ \mathrm{m/s^2}\right)}{50.0\ \mathrm{N/m}}}^{2} \quad .$$

8. Calculate d and convert units:
   Equation:

$$d = 1.59\ \mathrm{m}.$$

Discussion b
This is the total distance traveled back and forth across $x=0$, which is the undamped equilibrium position. The number of oscillations about the equilibrium position will be more than $d/X=(1.59\text{ m})/(0.100\text{ m})=15.9$ because the amplitude of the oscillations is decreasing with time. At the end of the motion, this system will not return to $x=0$ for this type of damping force, because static friction will exceed the restoring force. This system is underdamped. In contrast, an overdamped system with a simple constant damping force would not cross the equilibrium position $x=0$ a single time. For example, if this system had a damping force 20 times greater, it would only move 0.0484 m toward the equilibrium position from its original 0.100-m position.
This worked example illustrates how to apply problem-solving strategies to situations that integrate the different concepts you have learned. The first step is to identify the physical principles involved in the problem. The second step is to solve for the unknowns using familiar problem-solving strategies. These are found throughout the text, and many worked examples show how to use them for single topics. In this integrated concepts example, you can see how to apply them across several topics. You will find these techniques useful in applications of physics outside a physics course, such as in your profession, in other science disciplines, and in everyday life.

Exercise:
Check Your Understanding

Problem: Why are completely undamped harmonic oscillators so rare?

Solution:

Friction often comes into play whenever an object is moving. Friction causes damping in a harmonic oscillator.

Exercise:
Check Your Understanding

Problem: Describe the difference between overdamping, underdamping, and critical damping.

Solution:

An overdamped system moves slowly toward equilibrium. An underdamped system moves quickly to equilibrium, but will oscillate about the equilibrium point as it does so. A critically damped system moves as quickly as possible toward equilibrium without oscillating about the equilibrium.

Section Summary

- Damped harmonic oscillators have non-conservative forces that dissipate their energy.
- Critical damping returns the system to equilibrium as fast as possible without overshooting.
- An underdamped system will oscillate through the equilibrium position.
- An overdamped system moves more slowly toward equilibrium than one that is critically damped.

Conceptual Questions

Exercise:

Problem:

Give an example of a damped harmonic oscillator. (They are more common than undamped or simple harmonic oscillators.)

Exercise:

Problem: How would a car bounce after a bump under each of these conditions?

- overdamping
- underdamping
- critical damping

Exercise:

Problem:

Most harmonic oscillators are damped and, if undriven, eventually come to a stop. How is this observation related to the second law of thermodynamics?

Problems & Exercises

Exercise:

Problem:

The amplitude of a lightly damped oscillator decreases by 3.0% during each cycle. What percentage of the mechanical energy of the oscillator is lost in each cycle?

Glossary

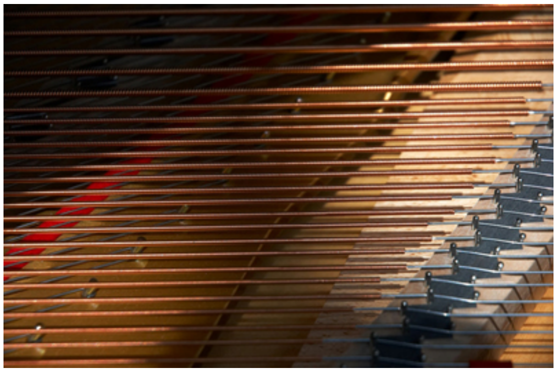
critical damping
: the condition in which the damping of an oscillator causes it to return as quickly as possible to its equilibrium position without oscillating back and forth about this position

over damping
: the condition in which damping of an oscillator causes it to return to equilibrium without oscillating; oscillator moves more slowly toward equilibrium than in the critically damped system

under damping
: the condition in which damping of an oscillator causes it to return to equilibrium with the amplitude gradually decreasing to zero; system returns to equilibrium faster but overshoots and crosses the equilibrium position one or more times

Forced Oscillations and Resonance

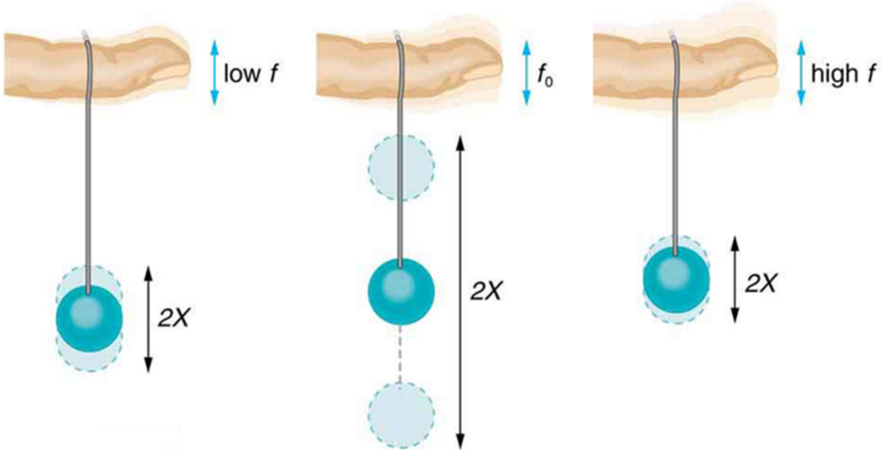
- Observe resonance of a paddle ball on a string.
- Observe amplitude of a damped harmonic oscillator.



You can cause the strings in a piano to vibrate simply by producing sound waves from your voice. (credit: Matt Billings, Flickr)

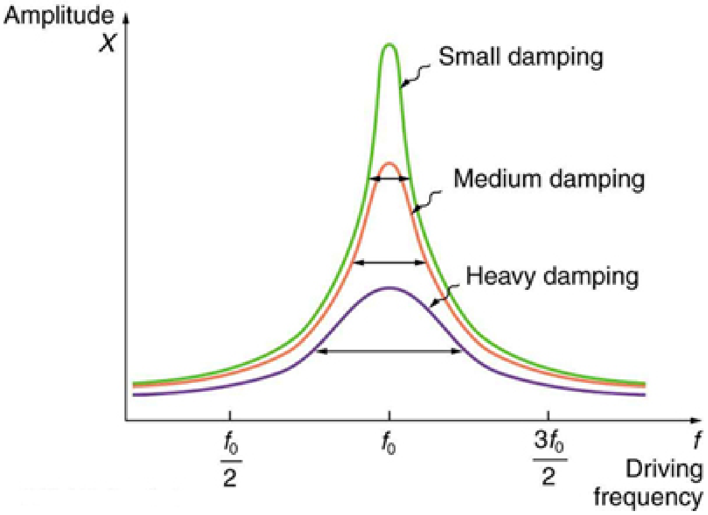
Sit in front of a piano sometime and sing a loud brief note at it with the dampers off its strings. It will sing the same note back at you—the strings, having the same frequencies as your voice, are resonating in response to the forces from the sound waves that you sent to them. Your voice and a piano's strings is a good example of the fact that objects—in this case, piano strings—can be forced to oscillate but oscillate best at their natural frequency. In this section, we shall briefly explore applying a *periodic driving force* acting on a simple harmonic oscillator. The driving force puts energy into the system at a certain frequency, not necessarily the same as the natural frequency of the system. The **natural frequency** is the frequency at which a system would oscillate if there were no driving and no damping force.

Most of us have played with toys involving an object supported on an elastic band, something like the paddle ball suspended from a finger in [link]. Imagine the finger in the figure is your finger. At first you hold your finger steady, and the ball bounces up and down with a small amount of damping. If you move your finger up and down slowly, the ball will follow along without bouncing much on its own. As you increase the frequency at which you move your finger up and down, the ball will respond by oscillating with increasing amplitude. When you drive the ball at its natural frequency, the ball's oscillations increase in amplitude with each oscillation for as long as you drive it. The phenomenon of driving a system with a frequency equal to its natural frequency is called **resonance**. A system being driven at its natural frequency is said to **resonate**. As the driving frequency gets progressively higher than the resonant or natural frequency, the amplitude of the oscillations becomes smaller, until the oscillations nearly disappear and your finger simply moves up and down with little effect on the ball.



The paddle ball on its rubber band moves in response to the finger supporting it. If the finger moves with the natural frequency f_0 of the ball on the rubber band, then a resonance is achieved, and the amplitude of the ball's oscillations increases dramatically. At higher and lower driving frequencies, energy is transferred to the ball less efficiently, and it responds with lower-amplitude oscillations.

[link] shows a graph of the amplitude of a damped harmonic oscillator as a function of the frequency of the periodic force driving it. There are three curves on the graph, each representing a different amount of damping. All three curves peak at the point where the frequency of the driving force equals the natural frequency of the harmonic oscillator. The highest peak, or greatest response, is for the least amount of damping, because less energy is removed by the damping force.

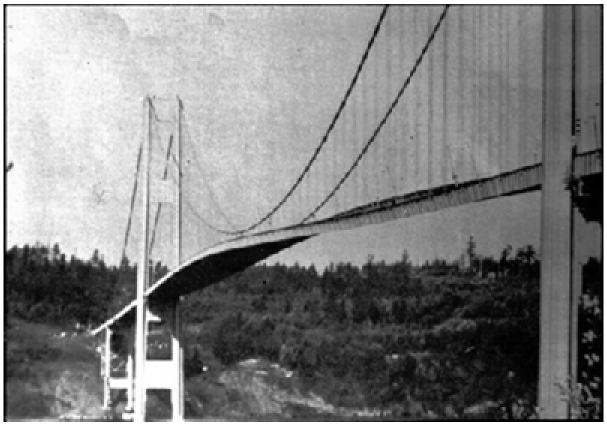


Amplitude of a harmonic oscillator as a function of the frequency of the driving force. The curves represent the same oscillator with the same natural frequency but with different amounts of damping. Resonance occurs when the driving frequency equals the natural frequency, and the greatest response is for the least amount of damping. The narrowest response is also for the least damping.

It is interesting that the widths of the resonance curves shown in [link] depend on damping: the less the damping, the narrower the resonance. The message is that if you want a driven oscillator to resonate at a very specific frequency, you need as little damping as possible. Little damping is the case for piano strings and many other musical instruments. Conversely, if you want small-amplitude oscillations, such as in a car’s suspension system, then you want heavy damping. Heavy damping reduces the amplitude, but the tradeoff is that the system responds at more frequencies.

These features of driven harmonic oscillators apply to a huge variety of systems. When you tune a radio, for example, you are adjusting its resonant frequency so that it only oscillates to the desired station’s broadcast (driving) frequency. The more selective the radio is in discriminating between stations, the smaller its damping. Magnetic resonance imaging (MRI) is a widely used medical diagnostic tool in which atomic nuclei (mostly hydrogen nuclei) are made to resonate by incoming radio waves (on the order of 100 MHz). A child on a swing is driven by a parent at the swing’s natural frequency to achieve maximum amplitude. In all of these cases, the efficiency of energy transfer from the driving force into the oscillator is best at resonance. Speed bumps and gravel roads prove that even a car’s suspension system is not immune to resonance. In spite of finely engineered shock absorbers, which ordinarily convert mechanical energy to thermal energy almost as fast as it comes in, speed bumps still cause a large-amplitude oscillation. On gravel roads that are corrugated, you may have noticed that if you travel at the “wrong” speed, the bumps are very noticeable whereas at other speeds you may hardly feel the bumps at all. [link] shows a photograph of a famous example (the Tacoma Narrows Bridge) of the destructive effects of a driven harmonic oscillation. The Millennium Bridge in London was closed for a short period of time for the same reason while inspections were carried out.

In our bodies, the chest cavity is a clear example of a system at resonance. The diaphragm and chest wall drive the oscillations of the chest cavity which result in the lungs inflating and deflating. The system is critically damped and the muscular diaphragm oscillates at the resonant value for the system, making it highly efficient.



In 1940, the Tacoma Narrows Bridge in Washington state collapsed. Heavy cross winds drove the bridge into oscillations at its resonant frequency. Damping decreased when support cables broke loose and started to slip over the towers, allowing increasingly greater amplitudes until the structure failed (credit: PRI's *Studio 360*, via Flickr)

Exercise:
Check Your Understanding

Problem:

A famous magic trick involves a performer singing a note toward a crystal glass until the glass shatters. Explain why the trick works in terms of resonance and natural frequency.

Solution:

The performer must be singing a note that corresponds to the natural frequency of the glass. As the sound wave is directed at the glass, the glass responds by resonating at the same frequency as the sound wave.

With enough energy introduced into the system, the glass begins to vibrate and eventually shatters.

Section Summary

- A system’s natural frequency is the frequency at which the system will oscillate if not affected by driving or damping forces.
- A periodic force driving a harmonic oscillator at its natural frequency produces resonance. The system is said to resonate.
- The less damping a system has, the higher the amplitude of the forced oscillations near resonance. The more damping a system has, the broader response it has to varying driving frequencies.

Conceptual Questions

Exercise:

Problem:

Why are soldiers in general ordered to “route step” (walk out of step) across a bridge?

Problems & Exercises

Exercise:

Problem:

How much energy must the shock absorbers of a 1200-kg car dissipate in order to damp a bounce that initially has a velocity of 0.800 m/s at the equilibrium position? Assume the car returns to its original vertical position.

Solution:

384 J

Exercise:

Problem:

If a car has a suspension system with a force constant of $5.00 \times 10^4 \text{ N/m}$, how much energy must the car's shocks remove to dampen an oscillation starting with a maximum displacement of 0.0750 m?

Exercise:

Problem:

(a) How much will a spring that has a force constant of 40.0 N/m be stretched by an object with a mass of 0.500 kg when hung motionless from the spring? (b) Calculate the decrease in gravitational potential energy of the 0.500-kg object when it descends this distance. (c) Part of this gravitational energy goes into the spring. Calculate the energy stored in the spring by this stretch, and compare it with the gravitational potential energy. Explain where the rest of the energy might go.

Solution:

(a). 0.123 m

(b). −0.600 J

(c). 0.300 J. The rest of the energy may go into heat caused by friction and other damping forces.

Exercise:

Problem:

Suppose you have a 0.750-kg object on a horizontal surface connected to a spring that has a force constant of 150 N/m. There is simple friction between the object and surface with a static coefficient of friction $\mu_{\text{s}} = 0.100$. (a) How far can the spring be stretched without moving the mass? (b) If the object is set into oscillation with an amplitude twice the distance found in part (a), and the kinetic coefficient of friction is $\mu_{\text{k}} = 0.0850$, what total distance does it travel before stopping? Assume it starts at the maximum amplitude.

Exercise:

Problem:

Engineering Application: A suspension bridge oscillates with an effective force constant of $1.00 \times 10^{8}\ \text{N/m}$. (a) How much energy is needed to make it oscillate with an amplitude of 0.100 m? (b) If soldiers march across the bridge with a cadence equal to the bridge’s natural frequency and impart $1.00 \times 10^{4}\ \text{J}$ of energy each second, how long does it take for the bridge’s oscillations to go from 0.100 m to 0.500 m amplitude?

Solution:

(a) $5.00 \times 10^{5}\ \text{J}$

(b) $1.20{\times}10^{3}\ \text{s}$

Glossary

natural frequency
: the frequency at which a system would oscillate if there were no driving and no damping forces

resonance

the phenomenon of driving a system with a frequency equal to the system's natural frequency

resonate

a system being driven at its natural frequency

Waves

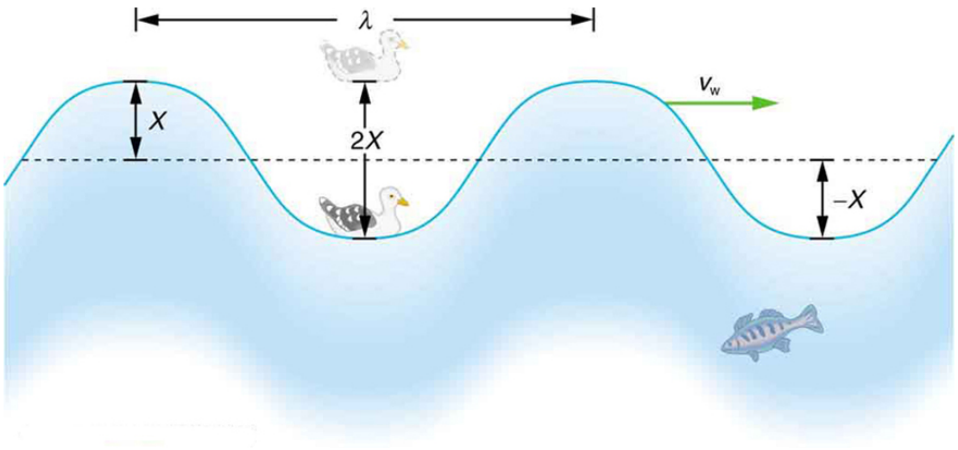
- State the characteristics of a wave.
- Calculate the velocity of wave propagation.

Waves in the ocean behave similarly to all other types of waves. (credit: Steve Jurveston, Flickr)

What do we mean when we say something is a wave? The most intuitive and easiest wave to imagine is the familiar water wave. More precisely, a **wave** is a disturbance that propagates, or moves from the place it was created. For water waves, the disturbance is in the surface of the water, perhaps created by a rock thrown into a pond or by a swimmer splashing the surface repeatedly. For sound waves, the disturbance is a change in air pressure, perhaps created by the oscillating cone inside a speaker. For earthquakes, there are several types of disturbances, including disturbance of Earth’s surface and pressure disturbances under the surface. Even radio waves are most easily understood using an analogy with water waves. Visualizing water waves is useful because there is more to it than just a mental image. Water waves exhibit characteristics common to all waves, such as amplitude, period, frequency and energy. All wave characteristics can be described by a small set of underlying principles.

A wave is a disturbance that propagates, or moves from the place it was created. The simplest waves repeat themselves for several cycles and are associated with simple harmonic motion. Let us start by considering the simplified water wave in [link]. The wave is an up and down disturbance of the water surface. It causes a sea gull to move up and down in simple harmonic motion as the wave crests and troughs (peaks and valleys) pass under the bird. The time for one complete up and down motion is the wave's period T. The wave's frequency is $f = 1/T$, as usual. The wave itself moves to the right in the figure. This movement of the wave is actually the disturbance moving to the right, not the water itself (or the bird would move to the right). We define **wave velocity** v_{w} to be the speed at which the disturbance moves. Wave velocity is sometimes also called the *propagation velocity or propagation speed*, because the disturbance propagates from one location to another.

Note:
Misconception Alert
Many people think that water waves push water from one direction to another. In fact, the particles of water tend to stay in one location, save for moving up and down due to the energy in the wave. The energy moves forward through the water, but the water stays in one place. If you feel yourself pushed in an ocean, what you feel is the energy of the wave, not a rush of water.



An idealized ocean wave passes under a sea gull that bobs up and down in simple harmonic motion. The wave has a wavelength λ, which is the distance between adjacent identical parts of the wave. The up and down disturbance of the surface propagates parallel to the surface at a speed v_{w}.

The water wave in the figure also has a length associated with it, called its **wavelength** λ, the distance between adjacent identical parts of a wave. (λ is the distance parallel to the direction of propagation.) The speed of propagation v_{w} is the distance the wave travels in a given time, which is one wavelength in the time of one period. In equation form, that is

Equation:

$$v_{\text{w}} = \frac{\lambda}{T}$$

or

Equation:

$$v_{\text{w}} = f\lambda.$$

This fundamental relationship holds for all types of waves. For water waves, v_{w} is the speed of a surface wave; for sound, v_{w} is the speed of sound; and for visible light, v_{w} is the speed of light, for example.

Note:

Take-Home Experiment: Waves in a Bowl

Fill a large bowl or basin with water and wait for the water to settle so there are no ripples. Gently drop a cork into the middle of the bowl. Estimate the wavelength and period of oscillation of the water wave that propagates away from the cork. Remove the cork from the bowl and wait for the water to settle again. Gently drop the cork at a height that is different from the first drop. Does the wavelength depend upon how high above the water the cork is dropped?

Example:
Calculate the Velocity of Wave Propagation: Gull in the Ocean

Calculate the wave velocity of the ocean wave in [link] if the distance between wave crests is 10.0 m and the time for a sea gull to bob up and down is 5.00 s.

Strategy

We are asked to find v_{w}. The given information tells us that $\lambda = 10.0\text{ m}$ and $T = 5.00\text{ s}$. Therefore, we can use $v_{\text{w}} = \frac{\lambda}{T}$ to find the wave velocity.

Solution

1. Enter the known values into $v_{\text{w}} = \frac{\lambda}{T}$:
   Equation:
   $$v_{\text{w}} = \frac{10.0\text{ m}}{5.00\text{ s}}.$$
2. Solve for v_{w} to find v_{w}= 2.00 m/s.

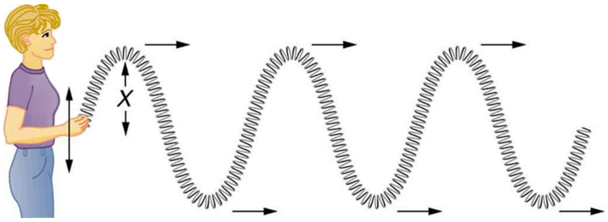
Discussion

This slow speed seems reasonable for an ocean wave. Note that the wave moves to the right in the figure at this speed, not the varying speed at which the sea gull moves up and down.

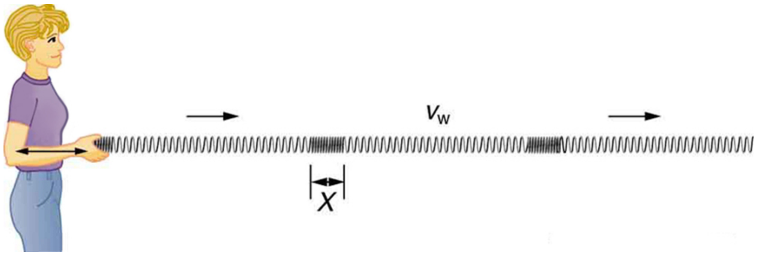
Transverse and Longitudinal Waves

A simple wave consists of a periodic disturbance that propagates from one place to another. The wave in [link] propagates in the horizontal direction while the surface is disturbed in the vertical direction. Such a wave is called a **transverse wave** or shear wave; in such a wave, the disturbance is perpendicular to the direction of propagation. In contrast, in a **longitudinal**

wave or compressional wave, the disturbance is parallel to the direction of propagation. [link] shows an example of a longitudinal wave. The size of the disturbance is its amplitude X and is completely independent of the speed of propagation v_w.



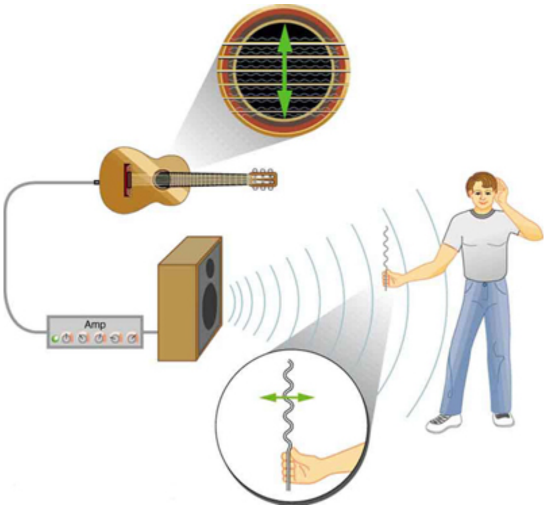
In this example of a transverse wave, the wave propagates horizontally, and the disturbance in the cord is in the vertical direction.



In this example of a longitudinal wave, the wave propagates horizontally, and the disturbance in the cord is also in the horizontal direction.

Waves may be transverse, longitudinal, or *a combination of the two*. (Water waves are actually a combination of transverse and longitudinal. The simplified water wave illustrated in [link] shows no longitudinal motion of the bird.) The waves on the strings of musical instruments are transverse—so are electromagnetic waves, such as visible light.

Sound waves in air and water are longitudinal. Their disturbances are periodic variations in pressure that are transmitted in fluids. Fluids do not have appreciable shear strength, and thus the sound waves in them must be longitudinal or compressional. Sound in solids can be both longitudinal and transverse.



The wave on a guitar string is transverse. The sound wave rattles a sheet of paper in a direction that shows the sound wave is longitudinal.

Earthquake waves under Earth’s surface also have both longitudinal and transverse components (called compressional or P-waves and shear or S-waves, respectively). These components have important individual characteristics—they propagate at different speeds, for example.

Earthquakes also have surface waves that are similar to surface waves on water.

Exercise:

Check Your Understanding

Problem:

Why is it important to differentiate between longitudinal and transverse waves?

Solution:

In the different types of waves, energy can propagate in a different direction relative to the motion of the wave. This is important to understand how different types of waves affect the materials around them.

Note:

PhET Explorations: Wave on a String

Watch a string vibrate in slow motion. Wiggle the end of the string and make waves, or adjust the frequency and amplitude of an oscillator. Adjust the damping and tension. The end can be fixed, loose, or open.

https://phet.colorado.edu/sims/html/wave-on-a-string/latest/wave-on-a-string_en.html

Section Summary

- A wave is a disturbance that moves from the point of creation with a wave velocity v_{w}.
- A wave has a wavelength λ, which is the distance between adjacent identical parts of the wave.
- Wave velocity and wavelength are related to the wave’s frequency and period by $v_{\text{w}} = \frac{\lambda}{T}$ or $v_{\text{w}} = f\lambda$.

- A transverse wave has a disturbance perpendicular to its direction of propagation, whereas a longitudinal wave has a disturbance parallel to its direction of propagation.

Conceptual Questions

Exercise:

Problem:

Give one example of a transverse wave and another of a longitudinal wave, being careful to note the relative directions of the disturbance and wave propagation in each.

Exercise:

Problem:

What is the difference between propagation speed and the frequency of a wave? Does one or both affect wavelength? If so, how?

Problems & Exercises

Exercise:

Problem:

Storms in the South Pacific can create waves that travel all the way to the California coast, which are 12,000 km away. How long does it take them if they travel at 15.0 m/s?

Solution:
Equation:

$$t=9.26\text{ d}$$

Exercise:

Problem:

Waves on a swimming pool propagate at 0.750 m/s. You splash the water at one end of the pool and observe the wave go to the opposite end, reflect, and return in 30.0 s. How far away is the other end of the pool?

Exercise:

Problem:

Wind gusts create ripples on the ocean that have a wavelength of 5.00 cm and propagate at 2.00 m/s. What is their frequency?

Solution:
Equation:

$$f = 40.0\ \text{Hz}$$

Exercise:

Problem:

How many times a minute does a boat bob up and down on ocean waves that have a wavelength of 40.0 m and a propagation speed of 5.00 m/s?

Exercise:

Problem:

Scouts at a camp shake the rope bridge they have just crossed and observe the wave crests to be 8.00 m apart. If they shake it the bridge twice per second, what is the propagation speed of the waves?

Solution:
Equation:

$$v_{\text{w}} = 16.0\ \text{m/s}$$

Exercise:

Problem:

What is the wavelength of the waves you create in a swimming pool if you splash your hand at a rate of 2.00 Hz and the waves propagate at 0.800 m/s?

Exercise:

Problem:

What is the wavelength of an earthquake that shakes you with a frequency of 10.0 Hz and gets to another city 84.0 km away in 12.0 s?

Solution:
Equation:

$$\lambda = 700\text{ m}$$

Exercise:

Problem:

Radio waves transmitted through space at $3.00 \times 10^{8}\text{ m/s}$ by the Voyager spacecraft have a wavelength of 0.120 m. What is their frequency?

Exercise:

Problem:

Your ear is capable of differentiating sounds that arrive at the ear just 1.00 ms apart. What is the minimum distance between two speakers that produce sounds that arrive at noticeably different times on a day when the speed of sound is 340 m/s?

Solution:
Equation:

$$d = 34.0\text{ cm}$$

Exercise:

Problem:

(a) Seismographs measure the arrival times of earthquakes with a precision of 0.100 s. To get the distance to the epicenter of the quake, they compare the arrival times of S- and P-waves, which travel at different speeds. [link]) If S- and P-waves travel at 4.00 and 7.20 km/s, respectively, in the region considered, how precisely can the distance to the source of the earthquake be determined? (b) Seismic waves from underground detonations of nuclear bombs can be used to locate the test site and detect violations of test bans. Discuss whether your answer to (a) implies a serious limit to such detection. (Note also that the uncertainty is greater if there is an uncertainty in the propagation speeds of the S- and P-waves.)

A seismograph as described in above problem.(credit: Oleg Alexandrov)

Glossary

longitudinal wave

a wave in which the disturbance is parallel to the direction of propagation

transverse wave
a wave in which the disturbance is perpendicular to the direction of propagation

wave velocity
the speed at which the disturbance moves. Also called the propagation velocity or propagation speed

wavelength
the distance between adjacent identical parts of a wave

Superposition and Interference

- Explain standing waves.
- Describe the mathematical representation of overtones and beat frequency.

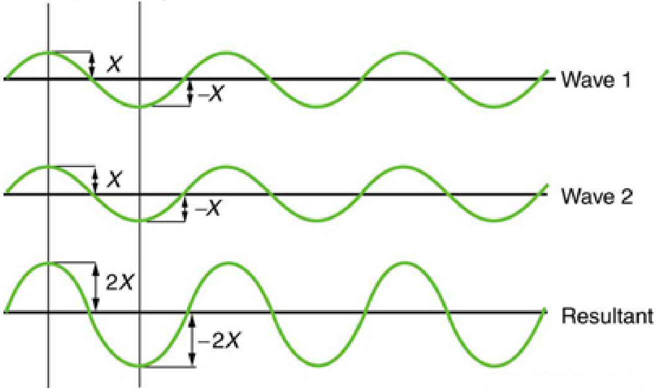
These waves result from the superposition of several waves from different sources, producing a complex pattern. (credit: waterborough, Wikimedia Commons)

Most waves do not look very simple. They look more like the waves in [link] than like the simple water wave considered in Waves. (Simple waves may be created by a simple harmonic oscillation, and thus have a sinusoidal shape). Complex waves are more interesting, even beautiful, but they look formidable. Most waves appear complex because they result from several simple waves adding together. Luckily, the rules for adding waves are quite simple.

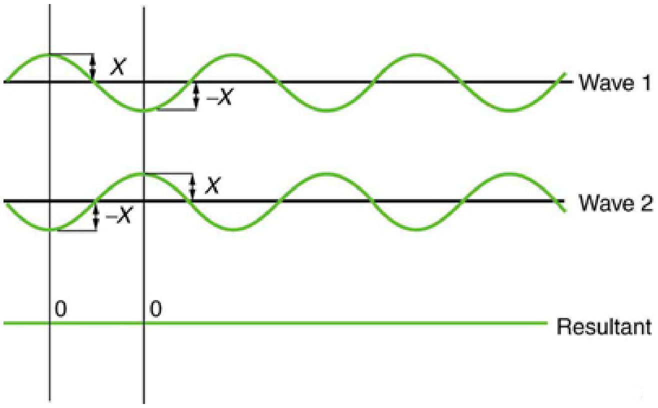
When two or more waves arrive at the same point, they superimpose themselves on one another. More specifically, the disturbances of waves are superimposed when they come together—a phenomenon called **superposition**. Each disturbance corresponds to a force, and forces add. If the disturbances are along the same line, then the resulting wave is a simple addition of the disturbances of the individual waves—that is, their amplitudes add. [link] and [link] illustrate superposition in two special cases, both of which produce simple results.

[link] shows two identical waves that arrive at the same point exactly in phase. The crests of the two waves are precisely aligned, as are the troughs. This superposition produces pure **constructive interference**. Because the disturbances add, pure constructive interference produces a wave that has twice the amplitude of the individual waves, but has the same wavelength.

[link] shows two identical waves that arrive exactly out of phase—that is, precisely aligned crest to trough—producing pure **destructive interference**. Because the disturbances are in the opposite direction for this superposition, the resulting amplitude is zero for pure destructive interference—the waves completely cancel.



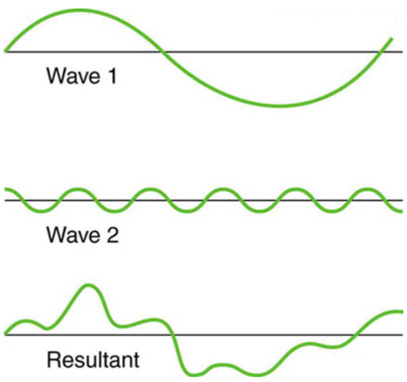
Pure constructive interference of two identical waves produces one with twice the amplitude, but the same wavelength.



Pure destructive interference of two identical waves produces zero amplitude, or complete cancellation.

While pure constructive and pure destructive interference do occur, they require precisely aligned identical waves. The superposition of most waves produces a combination of constructive and destructive interference and can vary from place to place and time to time. Sound from a stereo, for example, can be loud in one spot and quiet in another. Varying loudness means the sound waves add partially constructively and partially destructively at different locations. A stereo has at least two speakers creating sound waves, and waves can reflect from walls. All these waves superimpose. An example of sounds that vary over time from constructive to destructive is found in the combined whine of airplane jets heard by a stationary passenger. The combined sound can fluctuate up and down in volume as the sound from the two engines varies in time from constructive to destructive. These examples are of waves that are similar.

An example of the superposition of two dissimilar waves is shown in [link]. Here again, the disturbances add and subtract, producing a more complicated looking wave.

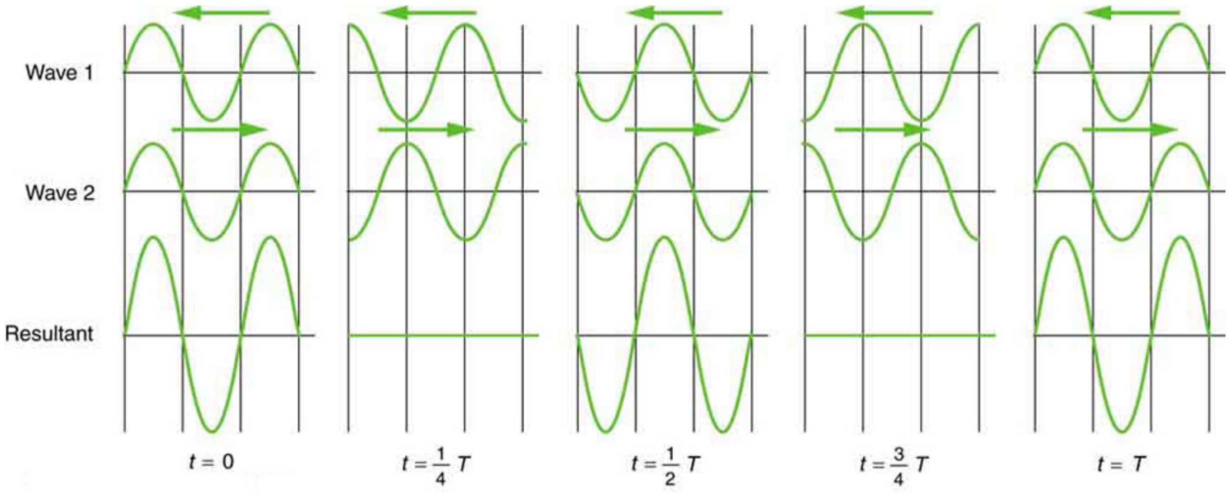


Superposition of non-identical waves exhibits both constructive and destructive interference.

Standing Waves

Sometimes waves do not seem to move; rather, they just vibrate in place. Unmoving waves can be seen on the surface of a glass of milk in a refrigerator, for example. Vibrations from the refrigerator motor create waves on the milk that oscillate up and down but do not seem to move across the surface. These waves are formed by the superposition of two or more moving waves, such as illustrated in [link] for two identical waves moving in opposite directions. The waves move through each other with their disturbances adding as they go by. If the two waves have the same amplitude and wavelength, then they alternate between constructive and destructive interference. The resultant looks like a wave standing in place and, thus, is called a **standing wave**. Waves on the glass of milk are one example of standing waves. There are other standing waves, such as on guitar strings and in organ pipes. With the glass of milk, the two waves that produce standing waves may come from reflections from the side of the glass.

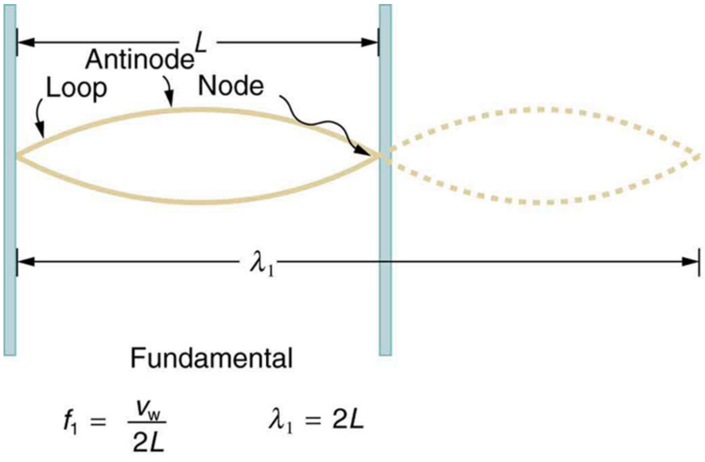
A closer look at earthquakes provides evidence for conditions appropriate for resonance, standing waves, and constructive and destructive interference. A building may be vibrated for several seconds with a driving frequency matching that of the natural frequency of vibration of the building—producing a resonance resulting in one building collapsing while neighboring buildings do not. Often buildings of a certain height are devastated while other taller buildings remain intact. The building height matches the condition for setting up a standing wave for that particular height. As the earthquake waves travel along the surface of Earth and reflect off denser rocks, constructive interference occurs at certain points. Often areas closer to the epicenter are not damaged while areas farther away are damaged.



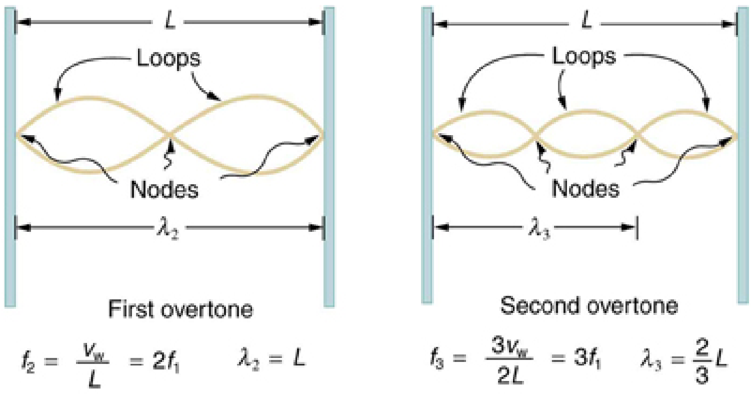
Standing wave created by the superposition of two identical waves moving in opposite directions. The oscillations are at fixed locations in space and result from alternately constructive and destructive interference.

Standing waves are also found on the strings of musical instruments and are due to reflections of waves from the ends of the string. [link] and [link] show three standing waves that can be created on a string that is fixed at both ends. **Nodes** are the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave. The fixed ends of strings must be nodes, too, because the string cannot move there. The word **antinode** is used to denote the location of maximum amplitude in standing waves. Standing waves on strings have a frequency that is related to the propagation speed v_{w} of the disturbance on the string. The wavelength λ is determined by the distance between the points where the string is fixed in place.

The lowest frequency, called the **fundamental frequency**, is thus for the longest wavelength, which is seen to be $\lambda_1 = 2L$. Therefore, the fundamental frequency is $f_1 = v_{\text{w}}/\lambda_1 = v_{\text{w}}/2L$. In this case, the **overtones** or harmonics are multiples of the fundamental frequency. As seen in [link], the first harmonic can easily be calculated since $\lambda_2 = L$. Thus, $f_2 = v_{\text{w}}/\lambda_2 = v_{\text{w}}/2L = 2f_1$. Similarly, $f_3 = 3f_1$, and so on. All of these frequencies can be changed by adjusting the tension in the string. The greater the tension, the greater v_{w} is and the higher the frequencies. This observation is familiar to anyone who has ever observed a string instrument being tuned. We will see in later chapters that standing waves are crucial to many resonance phenomena, such as in sounding boxes on string instruments.



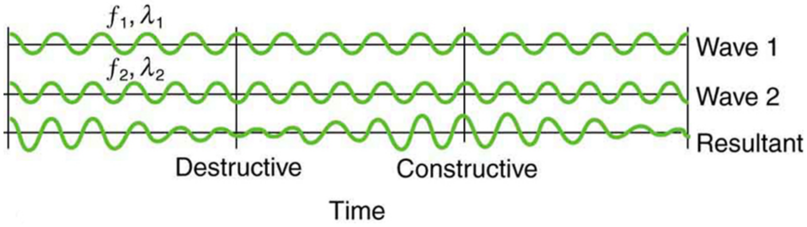
The figure shows a string oscillating at its fundamental frequency.



First and second harmonic frequencies are shown.

Beats

Striking two adjacent keys on a piano produces a warbling combination usually considered to be unpleasant. The superposition of two waves of similar but not identical frequencies is the culprit. Another example is often noticeable in jet aircraft, particularly the two-engine variety, while taxiing. The combined sound of the engines goes up and down in loudness. This varying loudness happens because the sound waves have similar but not identical frequencies. The discordant warbling of the piano and the fluctuating loudness of the jet engine noise are both due to alternately constructive and destructive interference as the two waves go in and out of phase. [link] illustrates this graphically.



Beats are produced by the superposition of two waves of slightly different frequencies but identical amplitudes. The waves alternate in time between constructive interference and destructive interference, giving the resulting wave a time-varying amplitude.

The wave resulting from the superposition of two similar-frequency waves has a frequency that is the average of the two. This wave fluctuates in amplitude, or *beats*, with a frequency called the **beat frequency**. We can determine the beat frequency by adding two waves together mathematically. Note that a wave can be represented at one point in space as

Equation:

$$x = X\cos\left(\frac{2\pi\, t}{T}\right) = X\cos(2\pi\ \mathrm{ft}),$$

where $f = 1/T$ is the frequency of the wave. Adding two waves that have different frequencies but identical amplitudes produces a resultant

Equation:

$$x = x_1 + x_2.$$

More specifically,

Equation:

$$x = X\cos(2\pi\ f_1 t) + X\cos(2\pi\ f_2 t).$$

Using a trigonometric identity, it can be shown that

Equation:

$$x = 2X\ \cos(\pi\ f_{\text{B}}t)\cos(2\pi\ f_{\text{ave}}t),$$

where

Equation:

$$f_{\text{B}} = \mid f_1 - f_2 \mid$$

is the beat frequency, and f_{ave} is the average of f_1 and f_2. These results mean that the resultant wave has twice the amplitude and the average frequency of the two superimposed waves, but it also fluctuates in overall amplitude at the beat frequency f_{B}. The first cosine term in the expression effectively causes the amplitude to go up and down. The second cosine term is the wave with frequency f_{ave}. This result is valid for all types of waves. However, if it is a sound wave, providing the two frequencies are similar, then what we hear is an average frequency that gets louder and softer (or warbles) at the beat frequency.

Note:

Making Career Connections

Piano tuners use beats routinely in their work. When comparing a note with a tuning fork, they listen for beats and adjust the string until the beats go away (to zero frequency). For example, if the tuning fork has a $256\ \text{Hz}$ frequency and two beats per second are heard, then the other frequency is either 254 or $258\ \text{Hz}$. Most keys hit multiple strings, and these strings are actually adjusted until they have nearly the same frequency and give a slow beat for richness. Twelve-string guitars and mandolins are also tuned using beats.

While beats may sometimes be annoying in audible sounds, we will find that beats have many applications. Observing beats is a very useful way to compare similar frequencies. There are applications of beats as apparently disparate as in ultrasonic imaging and radar speed traps.

Exercise:
Check Your Understanding

Problem:

Imagine you are holding one end of a jump rope, and your friend holds the other. If your friend holds her end still, you can move your end up and down, creating a transverse wave. If your friend then begins to move her end up and down, generating a wave in the opposite direction, what resultant wave forms would you expect to see in the jump rope?

Solution:

The rope would alternate between having waves with amplitudes two times the original amplitude and reaching equilibrium with no amplitude at all. The wavelengths will result in both constructive and destructive interference

Exercise:
Check Your Understanding

Problem: Define nodes and antinodes.

Solution:

Nodes are areas of wave interference where there is no motion. Antinodes are areas of wave interference where the motion is at its maximum point.

Exercise:
Check Your Understanding

Problem:

You hook up a stereo system. When you test the system, you notice that in one corner of the room, the sounds seem dull. In another area, the sounds seem excessively loud. Describe how the sound moving about the room could result in these effects.

Solution:

With multiple speakers putting out sounds into the room, and these sounds bouncing off walls, there is bound to be some wave interference. In the dull areas, the interference is probably mostly destructive. In the louder areas, the interference is probably mostly constructive.

Note:
PhET Explorations: Wave Interference
Make waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.

Wave Interference

Section Summary

- Superposition is the combination of two waves at the same location.
- Constructive interference occurs when two identical waves are superimposed in phase.

- Destructive interference occurs when two identical waves are superimposed exactly out of phase.
- A standing wave is one in which two waves superimpose to produce a wave that varies in amplitude but does not propagate.
- Nodes are points of no motion in standing waves.
- An antinode is the location of maximum amplitude of a standing wave.
- Waves on a string are resonant standing waves with a fundamental frequency and can occur at higher multiples of the fundamental, called overtones or harmonics.
- Beats occur when waves of similar frequencies f_1 and f_2 are superimposed. The resulting amplitude oscillates with a beat frequency given by
  Equation:

$$f_{\text{B}} = \mid f_1 - f_2 \mid .$$

Conceptual Questions

Exercise:

Problem:

Speakers in stereo systems have two color-coded terminals to indicate how to hook up the wires. If the wires are reversed, the speaker moves in a direction opposite that of a properly connected speaker. Explain why it is important to have both speakers connected the same way.

Problems & Exercises

Exercise:

Problem:

A car has two horns, one emitting a frequency of 199 Hz and the other emitting a frequency of 203 Hz. What beat frequency do they produce?

Solution:

$f = 4 \text{ Hz}$

Exercise:

Problem:

The middle-C hammer of a piano hits two strings, producing beats of 1.50 Hz. One of the strings is tuned to 260.00 Hz. What frequencies could the other string have?

Exercise:

Problem:

Two tuning forks having frequencies of 460 and 464 Hz are struck simultaneously. What average frequency will you hear, and what will the beat frequency be?

Solution:

462 Hz,

4 Hz

Exercise:

Problem:

Twin jet engines on an airplane are producing an average sound frequency of 4100 Hz with a beat frequency of 0.500 Hz. What are their individual frequencies?

Exercise:

Problem:

A wave traveling on a Slinky® that is stretched to 4 m takes 2.4 s to travel the length of the Slinky and back again. (a) What is the speed of the wave? (b) Using the same Slinky stretched to the same length, a standing wave is created which consists of three antinodes and four nodes. At what frequency must the Slinky be oscillating?

Solution:

(a) 3.33 m/s

(b) 1.25 Hz

Exercise:

Problem:

Three adjacent keys on a piano (F, F-sharp, and G) are struck simultaneously, producing frequencies of 349, 370, and 392 Hz. What beat frequencies are produced by this discordant combination?

Glossary

antinode
: the location of maximum amplitude in standing waves

beat frequency
: the frequency of the amplitude fluctuations of a wave

constructive interference
: when two waves arrive at the same point exactly in phase; that is, the crests of the two waves are precisely aligned, as are the troughs

destructive interference
: when two identical waves arrive at the same point exactly out of phase; that is, precisely aligned crest to trough

fundamental frequency
: the lowest frequency of a periodic waveform

nodes
: the points where the string does not move; more generally, nodes are where the wave disturbance is zero in a standing wave

overtones

multiples of the fundamental frequency of a sound

superposition

the phenomenon that occurs when two or more waves arrive at the same point

Energy in Waves: Intensity

- Calculate the intensity and the power of rays and waves.

The destructive effect of an earthquake is palpable evidence of the energy carried in these waves. The Richter scale rating of earthquakes is related to both their amplitude and the energy they carry. (credit: Petty Officer 2nd Class Candice Villarreal, U.S. Navy)

All waves carry energy. The energy of some waves can be directly observed. Earthquakes can shake whole cities to the ground, performing the work of thousands of wrecking balls.

Loud sounds pulverize nerve cells in the inner ear, causing permanent hearing loss. Ultrasound is used for deep-heat treatment of muscle strains. A laser beam can burn away a malignancy. Water waves chew up beaches.

The amount of energy in a wave is related to its amplitude. Large-amplitude earthquakes produce large ground displacements. Loud sounds have higher pressure amplitudes and come from larger-amplitude source vibrations than soft sounds. Large ocean breakers churn up the shore more than small ones. More quantitatively, a wave is a displacement that is resisted by a restoring force. The larger the displacement x, the larger the force $F = \text{kx}$ needed to create it. Because work W is related to force multiplied by distance (Fx) and energy is put into the wave by the work done to create it, the energy in a wave is related to amplitude. In fact, a wave's energy is directly proportional to its amplitude squared because

Equation:

$$W \propto \text{Fx} = \text{kx}^2.$$

The energy effects of a wave depend on time as well as amplitude. For example, the longer deep-heat ultrasound is applied, the more energy it transfers. Waves can also be concentrated or spread out. Sunlight, for example, can be focused to burn wood. Earthquakes spread out, so they do less damage the farther they get from the source. In both cases, changing the area the waves cover has important effects. All these pertinent factors are included in the definition of **intensity** I as power per unit area:

Equation:

$$I = \frac{P}{A}$$

where P is the power carried by the wave through area A. The definition of intensity is valid for any energy in transit, including that carried by waves. The SI unit for intensity is watts per square meter (W/m^2). For example, infrared and visible energy from the Sun impinge on Earth at an intensity of $1300\ \text{W/m}^2$ just above the atmosphere. There are other intensity-related units in use, too. The most common is the decibel. For example, a 90 decibel sound level corresponds to an intensity of $10^{-3}\ \text{W/m}^2$. (This quantity is not much power per unit area considering that 90 decibels is a relatively high sound level. Decibels will be discussed in some detail in a later chapter.

Example:

Calculating intensity and power: How much energy is in a ray of sunlight?

The average intensity of sunlight on Earth's surface is about $700\ \mathrm{W/m^2}$.
(a) Calculate the amount of energy that falls on a solar collector having an area of $0.500\ \mathrm{m^2}$ in 4.00 h.
(b) What intensity would such sunlight have if concentrated by a magnifying glass onto an area 200 times smaller than its own?

Strategy a

Because power is energy per unit time or $P = \frac{E}{t}$, the definition of intensity can be written as $I = \frac{P}{A} = \frac{E/t}{A}$, and this equation can be solved for E with the given information.

Solution a

1. Begin with the equation that states the definition of intensity:
   Equation:
   $$I = \frac{P}{A}.$$
2. Replace P with its equivalent E/t:
   Equation:
   $$I = \frac{E/t}{A}.$$
3. Solve for E:
   Equation:
   $$E = \mathrm{IAt}.$$
4. Substitute known values into the equation:
   Equation:
   $$E = \left(700\ \mathrm{W/m^2}\right)\left(0.500\ \mathrm{m^2}\right)\left[(4.00\ \mathrm{h})(3600\ \mathrm{s/h})\right].$$
5. Calculate to find E and convert units:

Equation:

$$5.04 \times 10^{6}\ \text{J},$$

Discussion a
The energy falling on the solar collector in 4 h in part is enough to be useful—for example, for heating a significant amount of water.
Strategy b
Taking a ratio of new intensity to old intensity and using primes for the new quantities, we will find that it depends on the ratio of the areas. All other quantities will cancel.
Solution b

1. Take the ratio of intensities, which yields:
   Equation:

$$\frac{I\prime}{I} = \frac{P\prime/A\prime}{P/A} = \frac{A}{A\prime}\left(\text{The powers cancel because } P\prime = P\right).$$

2. Identify the knowns:
   Equation:

$$A = 200A\prime,$$

   Equation:

$$\frac{I\prime}{I} = 200.$$

3. Substitute known quantities:
   Equation:

$$I\prime = 200I = 200\left(700\ \text{W/m}^{2}\right).$$

4. Calculate to find $I\prime$:
   Equation:

$$I\prime = 1.40 \times 10^5\ \text{W/m}^2.$$

Discussion b
Decreasing the area increases the intensity considerably. The intensity of the concentrated sunlight could even start a fire.

Example:
Determine the combined intensity of two waves: Perfect constructive interference
If two identical waves, each having an intensity of $1.00\ \text{W/m}^2$, interfere perfectly constructively, what is the intensity of the resulting wave?
Strategy
We know from Superposition and Interference that when two identical waves, which have equal amplitudes X, interfere perfectly constructively, the resulting wave has an amplitude of $2X$. Because a wave's intensity is proportional to amplitude squared, the intensity of the resulting wave is four times as great as in the individual waves.
Solution

1. Recall that intensity is proportional to amplitude squared.
2. Calculate the new amplitude:
   Equation:

$$I\prime \propto (X\prime)^2 = (2X)^2 = 4X^2.$$

3. Recall that the intensity of the old amplitude was:
   Equation:

$$I \propto X^2.$$

4. Take the ratio of new intensity to the old intensity. This gives:
   Equation:

$$\frac{I\prime}{I} = 4.$$

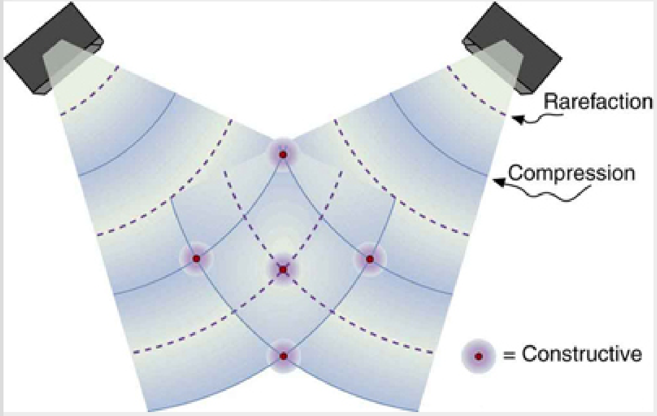
5. Calculate to find $I\prime$:

Equation:

$$I\prime = 4I = 4.00\ \text{W/m}^2.$$

Discussion

The intensity goes up by a factor of 4 when the amplitude doubles. This answer is a little disquieting. The two individual waves each have intensities of $1.00\ \text{W/m}^2$, yet their sum has an intensity of $4.00\ \text{W/m}^2$, which may appear to violate conservation of energy. This violation, of course, cannot happen. What does happen is intriguing. The area over which the intensity is $4.00\ \text{W/m}^2$ is much less than the area covered by the two waves before they interfered. There are other areas where the intensity is zero. The addition of waves is not as simple as our first look in Superposition and Interference suggested. We actually get a pattern of both constructive interference and destructive interference whenever two waves are added. For example, if we have two stereo speakers putting out $1.00\ \text{W/m}^2$ each, there will be places in the room where the intensity is $4.00\ \text{W/m}^2$, other places where the intensity is zero, and others in between. [link] shows what this interference might look like. We will pursue interference patterns elsewhere in this text.



These stereo speakers produce both constructive interference and destructive interference in the room, a property common to the superposition of all types of waves. The shading is proportional to intensity.

Exercise:
Check Your Understanding

Problem:

Which measurement of a wave is most important when determining the wave's intensity?

Solution:

Amplitude, because a wave’s energy is directly proportional to its amplitude squared.

Section Summary

Intensity is defined to be the power per unit area:

$I=\frac{P}{A}$ and has units of W/m^2.

Conceptual Questions

Exercise:

Problem:

Two identical waves undergo pure constructive interference. Is the resultant intensity twice that of the individual waves? Explain your answer.

Exercise:

Problem:

Circular water waves decrease in amplitude as they move away from where a rock is dropped. Explain why.

Problems & Exercises

Exercise:

Problem: **Medical Application**

Ultrasound of intensity $1.50 \times 10^{2}\ \text{W/m}^{2}$ is produced by the rectangular head of a medical imaging device measuring 3.00 by 5.00 cm. What is its power output?

Solution:

0.225 W

Exercise:

Problem:

The low-frequency speaker of a stereo set has a surface area of $0.05\ \text{m}^{2}$ and produces 1W of acoustical power. What is the intensity at the speaker? If the speaker projects sound uniformly in all directions, at what distance from the speaker is the intensity $0.1\ \text{W/m}^{2}$?

Exercise:

Problem:

To increase intensity of a wave by a factor of 50, by what factor should the amplitude be increased?

Solution:

7.07

Exercise:

Problem: **Engineering Application**

A device called an insolation meter is used to measure the intensity of sunlight has an area of 100 cm^2 and registers 6.50 W. What is the intensity in $\mathrm{W/m^2}$?

Exercise:

Problem: **Astronomy Application**

Energy from the Sun arrives at the top of the Earth’s atmosphere with an intensity of $1.30\ \mathrm{kW/m^2}$. How long does it take for $1.8 \times 10^9\ \mathrm{J}$ to arrive on an area of $1.00\ \mathrm{m^2}$?

Solution:

16.0 d

Exercise:

Problem:

Suppose you have a device that extracts energy from ocean breakers in direct proportion to their intensity. If the device produces 10.0 kW of power on a day when the breakers are 1.20 m high, how much will it produce when they are 0.600 m high?

Solution:

2.50 kW

Exercise:

Problem: **Engineering Application**

(a) A photovoltaic array of (solar cells) is 10.0% efficient in gathering solar energy and converting it to electricity. If the average intensity of sunlight on one day is $700\ \text{W/m}^2$, what area should your array have to gather energy at the rate of 100 W? (b) What is the maximum cost of the array if it must pay for itself in two years of operation averaging 10.0 hours per day? Assume that it earns money at the rate of 9.00 ¢ per kilowatt-hour.

Exercise:

Problem:

A microphone receiving a pure sound tone feeds an oscilloscope, producing a wave on its screen. If the sound intensity is originally $2.00 \times 10^{-5}\ \text{W/m}^2$, but is turned up until the amplitude increases by 30.0%, what is the new intensity?

Solution:

$3.38 \times 10^{-5}\ \text{W/m}^2$

Exercise:

Problem: Medical Application

(a) What is the intensity in W/m^2 of a laser beam used to burn away cancerous tissue that, when 90.0% absorbed, puts 500 J of energy into a circular spot 2.00 mm in diameter in 4.00 s? (b) Discuss how this intensity compares to the average intensity of sunlight (about $700\ \text{W/m}^2$) and the implications that would have if the laser beam entered your eye. Note how your answer depends on the time duration of the exposure.

Glossary

intensity
    power per unit area

Introduction to the Physics of Hearing class="introduction"

This tree fell some time ago. When it fell, atoms in the air were disturbed. Physicists would call this disturbance sound whether someone was around to hear it or not. (credit: B.A. Bowen Photography)

If a tree falls in the forest and no one is there to hear it, does it make a sound? The answer to this old philosophical question depends on how you define sound. If sound only exists when someone is around to perceive it, then there was no sound. However, if we define sound in terms of physics; that is, a disturbance of the atoms in matter transmitted from its origin outward (in other words, a wave), then there *was* a sound, even if nobody was around to hear it.

Such a wave is the physical phenomenon we call *sound*. Its perception is hearing. Both the physical phenomenon and its perception are interesting and will be considered in this text. We shall explore both sound and hearing; they are related, but are not the same thing. We will also explore the many practical uses of sound waves, such as in medical imaging.

Sound

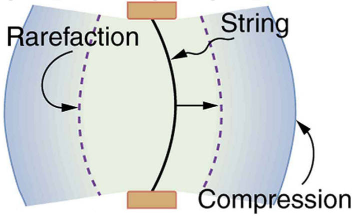
- Define sound and hearing.
- Describe sound as a longitudinal wave.

This glass has been shattered by a high-intensity sound wave of the same frequency as the resonant frequency of the glass. While the sound is not visible, the effects of the sound prove its existence. (credit: ||read||, Flickr)

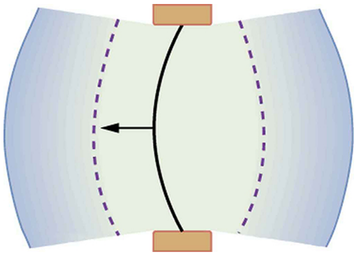
Sound can be used as a familiar illustration of waves. Because hearing is one of our most important senses, it is interesting to see how the physical properties of sound correspond to our perceptions of it. **Hearing** is the perception of sound, just as vision is the perception of visible light. But sound has important applications beyond hearing. Ultrasound, for example, is not heard but can be employed to form medical images and is also used in treatment.

The physical phenomenon of **sound** is defined to be a disturbance of matter that is transmitted from its source outward. Sound is a wave. On the atomic scale, it is a disturbance of atoms that is far more ordered than their thermal motions. In many instances, sound is a periodic wave, and the atoms undergo simple harmonic motion. In this text, we shall explore such periodic sound waves.

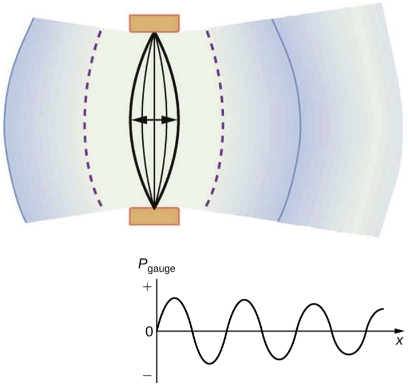
A vibrating string produces a sound wave as illustrated in [link], [link], and [link]. As the string oscillates back and forth, it transfers energy to the air, mostly as thermal energy created by turbulence. But a small part of the string's energy goes into compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions (high pressure regions) and rarefactions (low pressure regions) move out as longitudinal pressure waves having the same frequency as the string—they are the disturbance that is a sound wave. (Sound waves in air and most fluids are longitudinal, because fluids have almost no shear strength. In solids, sound waves can be both transverse and longitudinal.) [link] shows a graph of gauge pressure versus distance from the vibrating string.



A vibrating string moving to the right compresses the air in front of it and expands the air behind it.

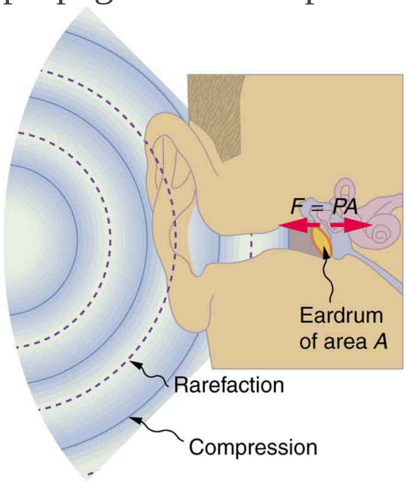


As the string moves to the left, it creates another compression and rarefaction as the ones on the right move away from the string.



After many vibrations, there are a series of compressions and rarefactions moving out from the string as a sound wave. The graph shows gauge pressure versus distance from the source. Pressures vary only slightly from atmospheric for ordinary sounds.

The amplitude of a sound wave decreases with distance from its source, because the energy of the wave is spread over a larger and larger area. But it is also absorbed by objects, such as the eardrum in [link], and converted to thermal energy by the viscosity of air. In addition, during each compression a little heat transfers to the air and during each rarefaction even less heat transfers from the air, so that the heat transfer reduces the organized disturbance into random thermal motions. (These processes can be viewed as a manifestation of the second law of thermodynamics presented in Introduction to the Second Law of Thermodynamics: Heat Engines and Their Efficiency.) Whether the heat transfer from compression to rarefaction is significant depends on how far apart they are—that is, it depends on wavelength. Wavelength, frequency, amplitude, and speed of propagation are important for sound, as they are for all waves.



Sound wave compressions and rarefactions travel up the ear canal and force the eardrum to vibrate. There is a net force on the eardrum, since the sound wave pressures differ from the atmospheric pressure found behind the eardrum. A complicated mechanism converts the vibrations to nerve impulses, which are perceived by the person.

Note:
PhET Explorations: Wave Interference
WMake waves with a dripping faucet, audio speaker, or laser! Add a second source or a pair of slits to create an interference pattern.
https://archive.cnx.org/specials/2fe7ad15-b00e-4402-b068-ff503985a18f/wave-interference/

Section Summary

- Sound is a disturbance of matter that is transmitted from its source outward.
- Sound is one type of wave.

- Hearing is the perception of sound.

Glossary

sound
: a disturbance of matter that is transmitted from its source outward

hearing
: the perception of sound

Speed of Sound, Frequency, and Wavelength

- Define pitch.
- Describe the relationship between the speed of sound, its frequency, and its wavelength.
- Describe the effects on the speed of sound as it travels through various media.
- Describe the effects of temperature on the speed of sound.

When a firework explodes, the light energy is perceived before the sound energy. Sound travels more slowly than light does. (credit: Dominic Alves, Flickr)

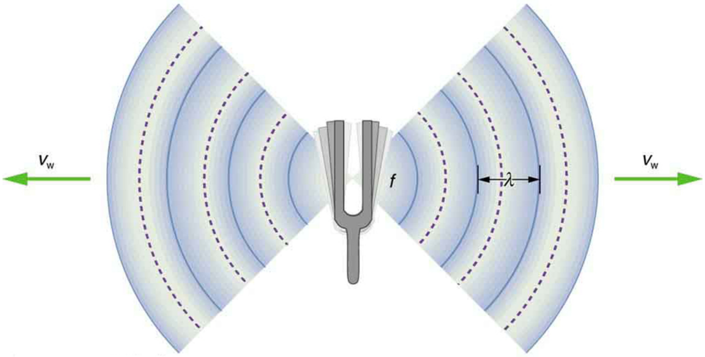
Sound, like all waves, travels at a certain speed and has the properties of frequency and wavelength. You can observe direct evidence of the speed of sound while watching a fireworks display. The flash of an explosion is seen well before its sound is heard, implying both that sound travels at a finite speed and that it is much slower than light. You can also directly sense the frequency of a sound. Perception of frequency is called **pitch**. The wavelength of sound is not directly sensed, but indirect evidence is found in the correlation of the size of musical instruments with their pitch. Small instruments, such as a piccolo, typically make high-pitch sounds, while large instruments, such as a tuba, typically make low-pitch sounds. High pitch means small wavelength, and the size of a musical instrument is directly related to the wavelengths of sound it produces. So a small instrument creates short-wavelength sounds. Similar arguments hold that a large instrument creates long-wavelength sounds.

The relationship of the speed of sound, its frequency, and wavelength is the same as for all waves:

Equation:

$$v_{\text{w}} = f\lambda,$$

where v_{w} is the speed of sound, f is its frequency, and λ is its wavelength. The wavelength of a sound is the distance between adjacent identical parts of a wave—for example, between adjacent compressions as illustrated in [link]. The frequency is the same as that of the source and is the number of waves that pass a point per unit time.



A sound wave emanates from a source vibrating at a frequency f, propagates at v_{w}, and has a wavelength λ.

[link] makes it apparent that the speed of sound varies greatly in different media. The speed of sound in a medium is determined by a combination of the medium’s rigidity (or compressibility in gases) and its density. The

more rigid (or less compressible) the medium, the faster the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is directly proportional to the stiffness of the oscillating object. The greater the density of a medium, the slower the speed of sound. This observation is analogous to the fact that the frequency of a simple harmonic motion is inversely proportional to the mass of the oscillating object. The speed of sound in air is low, because air is compressible. Because liquids and solids are relatively rigid and very difficult to compress, the speed of sound in such media is generally greater than in gases.

Medium	v_w**(m/s)**
Gases at 0^oC	
Air	331
Carbon dioxide	259
Oxygen	316
Helium	965
Hydrogen	1290
Liquids at 20^oC	
Ethanol	1160
Mercury	1450
Water, fresh	1480

Medium	**v_w(m/s)**
Sea water	1540
Human tissue	1540
Solids (longitudinal or bulk)	
Vulcanized rubber	54
Polyethylene	920
Marble	3810
Glass, Pyrex	5640
Lead	1960
Aluminum	5120
Steel	5960

Speed of Sound in Various Media

Earthquakes, essentially sound waves in Earth’s crust, are an interesting example of how the speed of sound depends on the rigidity of the medium. Earthquakes have both longitudinal and transverse components, and these travel at different speeds. The bulk modulus of granite is greater than its shear modulus. For that reason, the speed of longitudinal or pressure waves (P-waves) in earthquakes in granite is significantly higher than the speed of transverse or shear waves (S-waves). Both components of earthquakes travel slower in less rigid material, such as sediments. P-waves have speeds of 4 to 7 km/s, and S-waves correspondingly range in speed from 2 to 5 km/s, both being faster in more rigid material. The P-wave gets progressively farther ahead of the S-wave as they travel through Earth’s crust. The time between the P- and S-waves is routinely used to determine the distance to their source, the epicenter of the earthquake.

The speed of sound is affected by temperature in a given medium. For air at sea level, the speed of sound is given by

Equation:

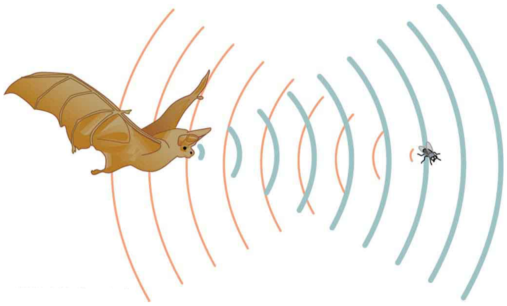
$$v_{\text{w}} = (331\ \text{m/s})\sqrt{\frac{T}{273\ \text{K}}},$$

where the temperature (denoted as T) is in units of kelvin. The speed of sound in gases is related to the average speed of particles in the gas, v_{rms}, and that

Equation:

$$v_{\text{rms}} = \sqrt{\frac{3\,kT}{m}},$$

where k is the Boltzmann constant (1.38×10^{-23} J/K) and m is the mass of each (identical) particle in the gas. So, it is reasonable that the speed of sound in air and other gases should depend on the square root of temperature. While not negligible, this is not a strong dependence. At 0ºC, the speed of sound is 331 m/s, whereas at 20.0ºC it is 343 m/s, less than a 4% increase. [link] shows a use of the speed of sound by a bat to sense distances. Echoes are also used in medical imaging.



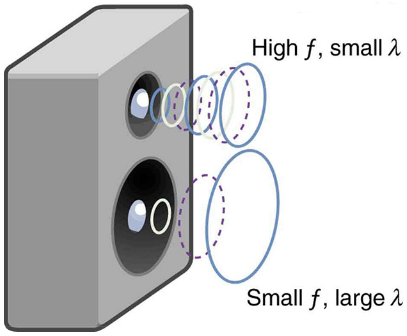
A bat uses sound echoes to find its way about and to catch prey. The time for the echo to return is directly proportional to the distance.

One of the more important properties of sound is that its speed is nearly independent of frequency. This independence is certainly true in open air for sounds in the audible range of 20 to 20,000 Hz. If this independence were not true, you would certainly notice it for music played by a marching band in a football stadium, for example. Suppose that high-frequency sounds traveled faster—then the farther you were from the band, the more the sound from the low-pitch instruments would lag that from the high-pitch ones. But the music from all instruments arrives in cadence independent of distance, and so all frequencies must travel at nearly the same speed. Recall that

Equation:

$$v_{\text{w}} = f\lambda.$$

In a given medium under fixed conditions, v_{w} is constant, so that there is a relationship between f and λ; the higher the frequency, the smaller the wavelength. See [link] and consider the following example.



Because they travel at the same speed in a given medium, low-frequency sounds must have a greater wavelength than high-frequency sounds.

Here, the lower-frequency sounds are emitted by the large speaker, called a woofer, while the higher-frequency sounds are emitted by the small speaker, called a tweeter.

Example:
Calculating Wavelengths: What Are the Wavelengths of Audible Sounds?

Calculate the wavelengths of sounds at the extremes of the audible range, 20 and 20,000 Hz, in 30.0ºC air. (Assume that the frequency values are accurate to two significant figures.)

Strategy

To find wavelength from frequency, we can use $v_{\text{w}} = f\lambda$.

Solution

1. Identify knowns. The value for v_{w}, is given by
   Equation:
   $$v_{\text{w}} = (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}}.$$
2. Convert the temperature into kelvin and then enter the temperature into the equation
   Equation:
   $$v_{\text{w}} = (331\text{ m/s})\sqrt{\frac{303\text{ K}}{273\text{ K}}} = 348.7\text{ m/s}.$$

3. Solve the relationship between speed and wavelength for λ:
   Equation:

$$\lambda = \frac{v_{\text{w}}}{f}.$$

4. Enter the speed and the minimum frequency to give the maximum wavelength:
   Equation:

$$\lambda_{\text{max}} = \frac{348.7\text{ m/s}}{20\text{ Hz}} = 17\text{ m}.$$

5. Enter the speed and the maximum frequency to give the minimum wavelength:
   Equation:

$$\lambda_{\text{min}} = \frac{348.7\text{ m/s}}{20{,}000\text{ Hz}} = 0.017\text{ m} = 1.7\text{ cm}.$$

Discussion
Because the product of f multiplied by λ equals a constant, the smaller f is, the larger λ must be, and vice versa.

The speed of sound can change when sound travels from one medium to another. However, the frequency usually remains the same because it is like a driven oscillation and has the frequency of the original source. If v_{w} changes and f remains the same, then the wavelength λ must change. That is, because $v_{\text{w}} = f\lambda$, the higher the speed of a sound, the greater its wavelength for a given frequency.

Note:
Making Connections: Take-Home Investigation—Voice as a Sound Wave

Suspend a sheet of paper so that the top edge of the paper is fixed and the bottom edge is free to move. You could tape the top edge of the paper to the edge of a table. Gently blow near the edge of the bottom of the sheet and note how the sheet moves. Speak softly and then louder such that the sounds hit the edge of the bottom of the paper, and note how the sheet moves. Explain the effects.

Exercise:
Check Your Understanding

Problem:

Imagine you observe two fireworks explode. You hear the explosion of one as soon as you see it. However, you see the other firework for several milliseconds before you hear the explosion. Explain why this is so.

Solution:

Sound and light both travel at definite speeds. The speed of sound is slower than the speed of light. The first firework is probably very close by, so the speed difference is not noticeable. The second firework is farther away, so the light arrives at your eyes noticeably sooner than the sound wave arrives at your ears.

Exercise:
Check Your Understanding

Problem:

You observe two musical instruments that you cannot identify. One plays high-pitch sounds and the other plays low-pitch sounds. How could you determine which is which without hearing either of them play?

Solution:

Compare their sizes. High-pitch instruments are generally smaller than low-pitch instruments because they generate a smaller wavelength.

Section Summary

The relationship of the speed of sound v_{w}, its frequency f, and its wavelength λ is given by

Equation:

$$v_{\text{w}} = f\lambda,$$

which is the same relationship given for all waves.

In air, the speed of sound is related to air temperature T by

Equation:

$$v_{\text{w}} = (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}}.$$

v_{w} is the same for all frequencies and wavelengths.

Conceptual Questions

Exercise:

Problem:

How do sound vibrations of atoms differ from thermal motion?

Exercise:

Problem:

When sound passes from one medium to another where its propagation speed is different, does its frequency or wavelength change? Explain your answer briefly.

Problems & Exercises

Exercise:

Problem:

When poked by a spear, an operatic soprano lets out a 1200-Hz shriek. What is its wavelength if the speed of sound is 345 m/s?

Solution:

0.288 m

Exercise:

Problem:

What frequency sound has a 0.10-m wavelength when the speed of sound is 340 m/s?

Exercise:

Problem:

Calculate the speed of sound on a day when a 1500 Hz frequency has a wavelength of 0.221 m.

Solution:

332 m/s

Exercise:

Problem:

(a) What is the speed of sound in a medium where a 100-kHz frequency produces a 5.96-cm wavelength? (b) Which substance in [link] is this likely to be?

Exercise:

Problem:

Show that the speed of sound in 20.0^{o}C air is 343 m/s, as claimed in the text.

Solution:
Equation:

$$\begin{aligned} v_{\text{w}} &= (331\text{ m/s})\sqrt{\frac{T}{273\text{ K}}} = (331\text{ m/s})\sqrt{\frac{293\text{ K}}{273\text{ K}}} \\ &= 343\text{ m/s} \end{aligned}$$

Exercise:

Problem:

Air temperature in the Sahara Desert can reach 56.0^{o}C (about 134^{o}F). What is the speed of sound in air at that temperature?

Exercise:

Problem:

Dolphins make sounds in air and water. What is the ratio of the wavelength of a sound in air to its wavelength in seawater? Assume air temperature is 20.0^{o}C.

Solution:

0.223

Exercise:

Problem:

A sonar echo returns to a submarine 1.20 s after being emitted. What is the distance to the object creating the echo? (Assume that the submarine is in the ocean, not in fresh water.)

Exercise:

Problem:

(a) If a submarine’s sonar can measure echo times with a precision of 0.0100 s, what is the smallest difference in distances it can detect? (Assume that the submarine is in the ocean, not in fresh water.)

(b) Discuss the limits this time resolution imposes on the ability of the sonar system to detect the size and shape of the object creating the echo.

Solution:

(a) 7.70 m

(b) This means that sonar is good for spotting and locating large objects, but it isn’t able to resolve smaller objects, or detect the detailed shapes of objects. Objects like ships or large pieces of airplanes can be found by sonar, while smaller pieces must be found by other means.

Exercise:

Problem:

A physicist at a fireworks display times the lag between seeing an explosion and hearing its sound, and finds it to be 0.400 s. (a) How far away is the explosion if air temperature is 24.0°C and if you neglect the time taken for light to reach the physicist? (b) Calculate the distance to the explosion taking the speed of light into account. Note that this distance is negligibly greater.

Exercise:

Problem:

Suppose a bat uses sound echoes to locate its insect prey, 3.00 m away. (See [link].) (a) Calculate the echo times for temperatures of 5.00°C and 35.0°C. (b) What percent uncertainty does this cause for the bat in locating the insect? (c) Discuss the significance of this uncertainty and whether it could cause difficulties for the bat. (In practice, the bat continues to use sound as it closes in, eliminating most of any difficulties imposed by this and other effects, such as motion of the prey.)

Solution:

(a) 18.0 ms, 17.1 ms

(b) 5.00%

(c) This uncertainty could definitely cause difficulties for the bat, if it didn’t continue to use sound as it closed in on its prey. A 5% uncertainty could be the difference between catching the prey around the neck or around the chest, which means that it could miss grabbing its prey.

Glossary

pitch
: the perception of the frequency of a sound

Sound Intensity and Sound Level

- Define intensity, sound intensity, and sound pressure level.
- Calculate sound intensity levels in decibels (dB).

Noise on crowded roadways like this one in Delhi makes it hard to hear others unless they shout. (credit: Lingaraj G J, Flickr)

In a quiet forest, you can sometimes hear a single leaf fall to the ground. After settling into bed, you may hear your blood pulsing through your ears. But when a passing motorist has his stereo turned up, you cannot even hear what the person next to you in your car is saying. We are all very familiar with the loudness of sounds and aware that they are related to how energetically the source is vibrating. In cartoons depicting a screaming person (or an animal making a loud noise), the cartoonist often shows an open mouth with a vibrating uvula, the hanging tissue at the back of the mouth, to suggest a loud sound coming from the throat [link]. High noise exposure is hazardous to hearing, and it is common for musicians to have hearing losses that are sufficiently severe that they interfere with the musicians’ abilities to perform. The relevant physical quantity is sound intensity, a concept that is valid for all sounds whether or not they are in the audible range.

Intensity is defined to be the power per unit area carried by a wave. Power is the rate at which energy is transferred by the wave. In equation form, **intensity** I is

Equation:

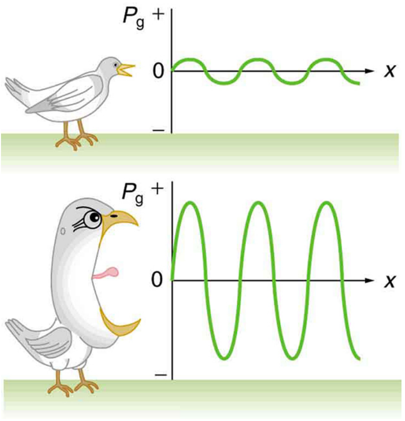
$$I=\frac{P}{A},$$

where P is the power through an area A. The SI unit for I is W/m^2. The intensity of a sound wave is related to its amplitude squared by the following relationship:

Equation:

$$I=\frac{(\Delta p)^2}{2\rho v_{\text{w}}}.$$

Here Δp is the pressure variation or pressure amplitude (half the difference between the maximum and minimum pressure in the sound wave) in units of pascals (Pa) or N/m^2. (We are using a lower case p for pressure to distinguish it from power, denoted by P above.) The energy (as kinetic energy $\frac{mv^2}{2}$) of an oscillating element of air due to a traveling sound wave is proportional to its amplitude squared. In this equation, ρ is the density of the material in which the sound wave travels, in units of kg/m^3, and v_{w} is the speed of sound in the medium, in units of m/s. The pressure variation is proportional to the amplitude of the oscillation, and so I varies as $(\Delta p)^2$ ([link]). This relationship is consistent with the fact that the sound wave is produced by some vibration; the greater its pressure amplitude, the more the air is compressed in the sound it creates.



Graphs of the gauge pressures in two sound waves of different intensities. The more intense sound is produced by a source that has larger-amplitude oscillations and has greater pressure maxima and minima. Because pressures are higher in the greater-intensity sound, it can exert larger forces on the objects it encounters.

Sound intensity levels are quoted in decibels (dB) much more often than sound intensities in watts per meter squared. Decibels are the unit of choice in the scientific literature as well as in the popular media. The reasons for this choice of units are related to how we perceive sounds. How our ears perceive sound can be more accurately described by the logarithm of the intensity rather than directly to the intensity. The **sound intensity level** β in decibels of a sound having an intensity I in watts per meter squared is defined to be

Equation:

$$\beta\ (\text{dB}) = 10\ \log_{10}\left(\frac{I}{I_0}\right),$$

where $I_0 = 10^{-12}\ \text{W/m}^2$ is a reference intensity. In particular, I_0 is the lowest or threshold intensity of sound a person with normal hearing can perceive at a frequency of 1000 Hz. Sound intensity level is not the same as intensity. Because β is defined in terms of a ratio, it is a unitless quantity telling you the *level* of the sound relative to a fixed standard ($10^{-12}\ \text{W/m}^2$, in this case). The units of decibels (dB) are used to indicate this ratio is multiplied by 10 in its definition. The bel, upon which the decibel is based, is named for Alexander Graham Bell, the inventor of the telephone.

Sound intensity level β (dB)	**Intensity $I(\text{W/m}^2)$**	**Example/effect**
0	1×10^{-12}	Threshold of hearing at 1000 Hz
10	1×10^{-11}	Rustle of leaves
20	1×10^{-10}	Whisper at 1 m distance
30	1×10^{-9}	Quiet home

Sound intensity level β (dB)	Intensity $I(\text{W/m}^2)$	Example/effect
40	1×10^{-8}	Average home
50	1×10^{-7}	Average office, soft music
60	1×10^{-6}	Normal conversation
70	1×10^{-5}	Noisy office, busy traffic
80	1×10^{-4}	Loud radio, classroom lecture
90	1×10^{-3}	Inside a heavy truck; damage from prolonged exposure[footnote] Several government agencies and health-related professional associations recommend that 85 dB not be exceeded for 8-hour daily exposures in the absence of hearing protection.
100	1×10^{-2}	Noisy factory, siren at 30 m; damage from 8 h per day exposure
110	1×10^{-1}	Damage from 30 min per day exposure
120	1	Loud rock concert, pneumatic chipper at 2 m; threshold of pain
140	1×10^{2}	Jet airplane at 30 m; severe pain, damage in seconds
160	1×10^{4}	Bursting of eardrums

Sound Intensity Levels and Intensities

Sound Intensity Levels and Intensities

The decibel level of a sound having the threshold intensity of $10^{–12}\ \text{W/m}^2$ is $\beta = 0\ \text{dB}$, because $\log_{10}1 = 0$. That is, the threshold of hearing is 0 decibels. [link] gives levels in decibels and intensities in watts per meter squared for some familiar sounds.

One of the more striking things about the intensities in [link] is that the intensity in watts per meter squared is quite small for most sounds. The ear is sensitive to as little as a trillionth of a watt per meter squared—even more impressive when you realize that the area of the eardrum is only about $1\ \text{cm}^2$, so that only $10^{–16}$ W falls on it at the threshold of hearing! Air molecules in a sound wave of this intensity vibrate over a distance of less than one molecular diameter, and the gauge pressures involved are less than $10^{–9}$ atm.

Another impressive feature of the sounds in [link] is their numerical range. Sound intensity varies by a factor of 10^{12} from threshold to a sound that causes damage in seconds. You are unaware of this tremendous range in sound intensity because how your ears respond can be described approximately as the logarithm of intensity. Thus, sound intensity levels in decibels fit your experience better than intensities in watts per meter squared. The decibel scale is also easier to relate to because most people are more accustomed to dealing with numbers such as 0, 53, or 120 than numbers such as $1.00 \times 10^{–11}$.

One more observation readily verified by examining [link] or using $I = \frac{(\Delta p)}{2\rho v_{\text{w}}}^{2}$ is that each factor of 10 in intensity corresponds to 10 dB. For example, a 90 dB sound compared with a 60 dB sound is 30 dB greater, or three factors of 10 (that is, 10^3 times) as intense. Another example is that if one sound is 10^7 as intense as another, it is 70 dB higher. See [link].

I_2/I_1	$\beta_2 - \beta_1$
2.0	3.0 dB
5.0	7.0 dB
10.0	10.0 dB

Ratios of Intensities and Corresponding Differences in Sound Intensity Levels

Example:
Calculating Sound Intensity Levels: Sound Waves

Calculate the sound intensity level in decibels for a sound wave traveling in air at 0ºC and having a pressure amplitude of 0.656 Pa.

Strategy

We are given Δp, so we can calculate I using the equation $I = (\Delta p)^2 / (2 p v_{\text{w}})^2$. Using I, we can calculate β straight from its definition in $\beta\ (\text{dB}) = 10\ \log_{10}(I/I_0)$.

Solution

(1) Identify knowns:

Sound travels at 331 m/s in air at 0ºC.

Air has a density of $1.29\ \text{kg/m}^3$ at atmospheric pressure and 0ºC.

(2) Enter these values and the pressure amplitude into $I = (\Delta p)^2 /\ (2 \rho v_{\text{w}})$:

Equation:

$$I = \frac{(\Delta p)^2}{2\rho v_{\text{w}}} = \frac{(0.656\ \text{Pa})^2}{2\left(1.29\ \text{kg/m}^3\right)(331\ \text{m/s})} = 5.04 \times 10^{-4}\ \text{W/m}^2.$$

(3) Enter the value for I and the known value for I_0 into $\beta\ (\text{dB}) = 10\ \log_{10}(I/I_0)$. Calculate to find the sound intensity level in decibels:

Equation:

$$10\ \log_{10}\left(5.04 \times 10^{8}\right) = 10\ \left(8.70\right)\ \text{dB} = 87\ \text{dB}.$$

Discussion

This 87 dB sound has an intensity five times as great as an 80 dB sound. So a factor of five in intensity corresponds to a difference of 7 dB in sound intensity level. This value is true for any intensities differing by a factor of five.

Example:

Change Intensity Levels of a Sound: What Happens to the Decibel Level?

Show that if one sound is twice as intense as another, it has a sound level about 3 dB higher.

Strategy

You are given that the ratio of two intensities is 2 to 1, and are then asked to find the difference in their sound levels in decibels. You can solve this problem using of the properties of logarithms.

Solution

(1) Identify knowns:

The ratio of the two intensities is 2 to 1, or:

Equation:

$$\frac{I_2}{I_1} = 2.00.$$

We wish to show that the difference in sound levels is about 3 dB. That is, we want to show:

Equation:

$$\beta_2 - \beta_1 = 3\ \text{dB}.$$

Note that:

Equation:

$$\log_{10} b - \log_{10} a = \log_{10}\left(\frac{b}{a}\right).$$

(2) Use the definition of β to get:
Equation:

$$\beta_2 - \beta_1 = 10\ \log_{10}\left(\frac{I_2}{I_1}\right) = 10\ \log_{10}2.00 = 10\ (0.301)\ \text{dB}.$$

Thus,
Equation:

$$\beta_2 - \beta_1 = 3.01\ \text{dB}.$$

Discussion
This means that the two sound intensity levels differ by 3.01 dB, or about 3 dB, as advertised. Note that because only the ratio I_2/I_1 is given (and not the actual intensities), this result is true for any intensities that differ by a factor of two. For example, a 56.0 dB sound is twice as intense as a 53.0 dB sound, a 97.0 dB sound is half as intense as a 100 dB sound, and so on.

It should be noted at this point that there is another decibel scale in use, called the **sound pressure level**, based on the ratio of the pressure amplitude to a reference pressure. This scale is used particularly in applications where sound travels in water. It is beyond the scope of most introductory texts to treat this scale because it is not commonly used for sounds in air, but it is important to note that very different decibel levels may be encountered when sound pressure levels are quoted. For example, ocean noise pollution produced by ships may be as great as 200 dB expressed in the sound pressure level, where the more familiar sound intensity level we use here would be something under 140 dB for the same sound.

Note:
Take-Home Investigation: Feeling Sound

Find a CD player and a CD that has rock music. Place the player on a light table, insert the CD into the player, and start playing the CD. Place your hand gently on the table next to the speakers. Increase the volume and note the level when the table just begins to vibrate as the rock music plays. Increase the reading on the volume control until it doubles. What has happened to the vibrations?

Exercise:
Check Your Understanding

Problem:

Describe how amplitude is related to the loudness of a sound.

Solution:

Amplitude is directly proportional to the experience of loudness. As amplitude increases, loudness increases.

Exercise:
Check Your Understanding

Problem:

Identify common sounds at the levels of 10 dB, 50 dB, and 100 dB.

Solution:

10 dB: Running fingers through your hair.

50 dB: Inside a quiet home with no television or radio.

100 dB: Take-off of a jet plane.

Section Summary

- Intensity is the same for a sound wave as was defined for all waves; it is
  Equation:

  $$I = \frac{P}{A},$$

  where P is the power crossing area A. The SI unit for I is watts per meter squared. The intensity of a sound wave is also related to the pressure amplitude Δp
  Equation:

  $$I = \frac{(\Delta p)^2}{2\rho v_{\mathrm{w}}},$$

  where ρ is the density of the medium in which the sound wave travels and v_{w} is the speed of sound in the medium.
- Sound intensity level in units of decibels (dB) is
  Equation:

  $$\beta\,(\mathrm{dB}) = 10\,\log_{10}\left(\frac{I}{I_0}\right),$$

  where $I_0 = 10^{-12}\ \mathrm{W/m^2}$ is the threshold intensity of hearing.

Conceptual Questions

Exercise:

Problem:

Six members of a synchronized swim team wear earplugs to protect themselves against water pressure at depths, but they can still hear the music and perform the combinations in the water perfectly. One day, they were asked to leave the pool so the dive team could practice a few dives, and they tried to practice on a mat, but seemed to have a lot more difficulty. Why might this be?

Exercise:

Problem:

A community is concerned about a plan to bring train service to their downtown from the town's outskirts. The current sound intensity level, even though the rail yard is blocks away, is 70 dB downtown. The mayor assures the public that there will be a difference of only 30 dB in sound in the downtown area. Should the townspeople be concerned? Why?

Problems & Exercises

Exercise:

Problem:

What is the intensity in watts per meter squared of 85.0-dB sound?

Solution:
Equation:

$$3.16 \times 10^{-4}\ \text{W/m}^2$$

Exercise:

Problem:

The warning tag on a lawn mower states that it produces noise at a level of 91.0 dB. What is this in watts per meter squared?

Exercise:

Problem:

A sound wave traveling in 20^{o}C air has a pressure amplitude of 0.5 Pa. What is the intensity of the wave?

Solution:
Equation:

$$3.04\times 10^{–4}\ \text{W/m}^{2}$$

Exercise:

Problem:

What intensity level does the sound in the preceding problem correspond to?

Exercise:

Problem:

What sound intensity level in dB is produced by earphones that create an intensity of $4.00\times 10^{-2}\ \text{W/m}^{2}$?

Solution:

106 dB

Exercise:

Problem:

Show that an intensity of $10^{–12}\ \text{W/m}^{2}$ is the same as $10^{–16}\ \text{W/cm}^{2}$.

Exercise:

Problem:

(a) What is the decibel level of a sound that is twice as intense as a 90.0-dB sound? (b) What is the decibel level of a sound that is one-fifth as intense as a 90.0-dB sound?

Solution:

(a) 93 dB

(b) 83 dB

Exercise:

Problem:

(a) What is the intensity of a sound that has a level 7.00 dB lower than a $4.00 \times 10^{-9}\ \mathrm{W/m^2}$ sound? (b) What is the intensity of a sound that is 3.00 dB higher than a $4.00 \times 10^{-9}\ \mathrm{W/m^2}$ sound?

Exercise:

Problem:

(a) How much more intense is a sound that has a level 17.0 dB higher than another? (b) If one sound has a level 23.0 dB less than another, what is the ratio of their intensities?

Solution:

(a) 50.1

(b) 5.01×10^{-3} or $\frac{1}{200}$

Exercise:

Problem:

People with good hearing can perceive sounds as low in level as –8.00 dB at a frequency of 3000 Hz. What is the intensity of this sound in watts per meter squared?

Exercise:

Problem:

If a large housefly 3.0 m away from you makes a noise of 40.0 dB, what is the noise level of 1000 flies at that distance, assuming interference has a negligible effect?

Solution:

70.0 dB

Exercise:

Problem:

Ten cars in a circle at a boom box competition produce a 120-dB sound intensity level at the center of the circle. What is the average sound intensity level produced there by each stereo, assuming interference effects can be neglected?

Exercise:

Problem:

The amplitude of a sound wave is measured in terms of its maximum gauge pressure. By what factor does the amplitude of a sound wave increase if the sound intensity level goes up by 40.0 dB?

Solution:

100

Exercise:

Problem:

If a sound intensity level of 0 dB at 1000 Hz corresponds to a maximum gauge pressure (sound amplitude) of $10^{-9}\ \text{atm}$, what is the maximum gauge pressure in a 60-dB sound? What is the maximum gauge pressure in a 120-dB sound?

Exercise:

Problem:

An 8-hour exposure to a sound intensity level of 90.0 dB may cause hearing damage. What energy in joules falls on a 0.800-cm-diameter eardrum so exposed?

Solution:
Equation:

$$1.45 \times 10^{-3}\ \text{J}$$

Exercise:

Problem:

(a) Ear trumpets were never very common, but they did aid people with hearing losses by gathering sound over a large area and concentrating it on the smaller area of the eardrum. What decibel increase does an ear trumpet produce if its sound gathering area is $900\ \text{cm}^2$ and the area of the eardrum is $0.500\ \text{cm}^2$, but the trumpet only has an efficiency of 5.00% in transmitting the sound to the eardrum? (b) Comment on the usefulness of the decibel increase found in part (a).

Exercise:

Problem:

Sound is more effectively transmitted into a stethoscope by direct contact than through the air, and it is further intensified by being concentrated on the smaller area of the eardrum. It is reasonable to assume that sound is transmitted into a stethoscope 100 times as effectively compared with transmission though the air. What, then, is the gain in decibels produced by a stethoscope that has a sound gathering area of $15.0\ \text{cm}^2$, and concentrates the sound onto two eardrums with a total area of $0.900\ \text{cm}^2$ with an efficiency of 40.0%?

Solution:

28.2 dB

Exercise:

Problem:

Loudspeakers can produce intense sounds with surprisingly small energy input in spite of their low efficiencies. Calculate the power input needed to produce a 90.0-dB sound intensity level for a 12.0-cm-diameter speaker that has an efficiency of 1.00%. (This value is the sound intensity level right at the speaker.)

Glossary

intensity
: the power per unit area carried by a wave

sound intensity level
: a unitless quantity telling you the level of the sound relative to a fixed standard

sound pressure level
: the ratio of the pressure amplitude to a reference pressure

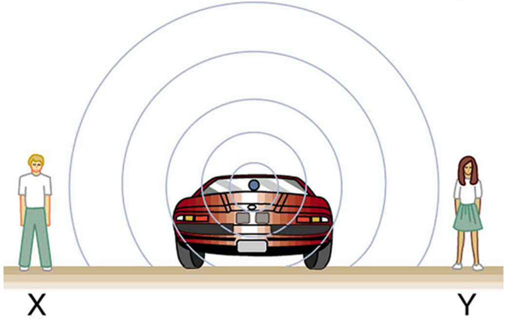
Doppler Effect and Sonic Booms

- Define Doppler effect, Doppler shift, and sonic boom.
- Calculate the frequency of a sound heard by someone observing Doppler shift.
- Describe the sounds produced by objects moving faster than the speed of sound.

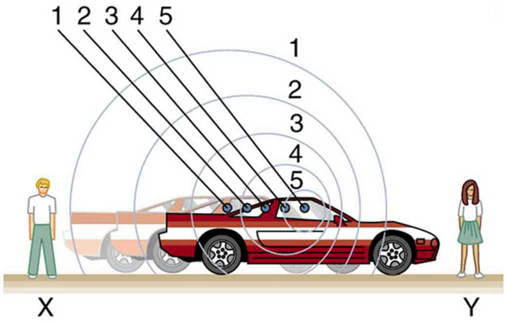
The characteristic sound of a motorcycle buzzing by is an example of the **Doppler effect**. The high-pitch scream shifts dramatically to a lower-pitch roar as the motorcycle passes by a stationary observer. The closer the motorcycle brushes by, the more abrupt the shift. The faster the motorcycle moves, the greater the shift. We also hear this characteristic shift in frequency for passing race cars, airplanes, and trains. It is so familiar that it is used to imply motion and children often mimic it in play.

The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer. Although less familiar, this effect is easily noticed for a stationary source and moving observer. For example, if you ride a train past a stationary warning bell, you will hear the bell’s frequency shift from high to low as you pass by. The actual change in frequency due to relative motion of source and observer is called a **Doppler shift**. The Doppler effect and Doppler shift are named for the Austrian physicist and mathematician Christian Johann Doppler (1803–1853), who did experiments with both moving sources and moving observers. Doppler, for example, had musicians play on a moving open train car and also play standing next to the train tracks as a train passed by. Their music was observed both on and off the train, and changes in frequency were measured.

What causes the Doppler shift? [link], [link], and [link] compare sound waves emitted by stationary and moving sources in a stationary air mass. Each disturbance spreads out spherically from the point where the sound was emitted. If the source is stationary, then all of the spheres representing the air compressions in the sound wave centered on the same point, and the stationary observers on either side see the same wavelength and frequency as emitted by the source, as in [link]. If the source is moving, as in [link], then the situation is different. Each compression of the air moves out in a sphere from the point where it was emitted, but the point of emission moves. This moving emission point causes the air compressions to be closer together on one side and farther apart on the other. Thus, the wavelength is shorter in the direction the source is moving (on the right in [link]), and longer in the opposite direction (on the left in [link]). Finally, if the observers move, as in [link], the frequency at which they receive the compressions changes. The observer moving toward the source receives them at a higher frequency, and the person moving away from the source receives them at a lower frequency.

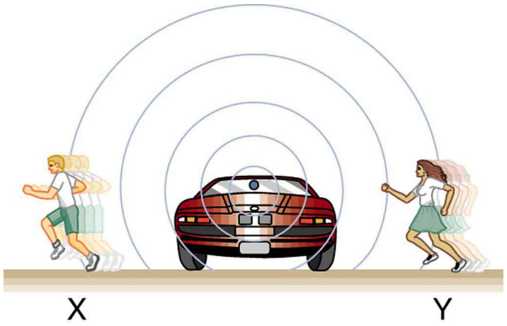


Sounds emitted by a source spread out in spherical waves. Because the source, observers, and air are stationary, the wavelength and frequency are the same in all directions and to all observers.



Sounds emitted by a

source moving to the right spread out from the points at which they were emitted. The wavelength is reduced and, consequently, the frequency is increased in the direction of motion, so that the observer on the right hears a higher-pitch sound. The opposite is true for the observer on the left, where the wavelength is increased and the frequency is reduced.



The same effect is produced when the observers move relative to the source. Motion toward the source increases frequency as the observer on the right passes through more wave crests than she would if stationary. Motion away from the source decreases frequency as the observer on the left passes through fewer wave crests than he would if stationary.

We know that wavelength and frequency are related by $v_{\text{w}} = f\lambda$, where v_{w} is the fixed speed of sound. The sound moves in a medium and has the same speed v_{w} in that medium whether the source is moving or not. Thus f multiplied by λ is a constant. Because the observer on the right in [link] receives a shorter wavelength, the frequency she receives must be higher. Similarly, the observer on the left receives a longer wavelength, and hence he hears a lower frequency. The same thing happens in [link]. A higher frequency is received by the observer moving toward the source, and a lower frequency is received by an observer moving away from the source. In general, then, relative motion of source and observer toward one another increases the received frequency. Relative motion apart decreases frequency. The greater the relative speed is, the greater the effect.

Note:

The Doppler Effect

The Doppler effect occurs not only for sound but for any wave when there is relative motion between the observer and the source. There are Doppler shifts in the frequency of sound, light, and water waves, for example. Doppler shifts can be used to determine velocity, such as when ultrasound is reflected from blood in a medical diagnostic. The recession of galaxies is determined by the shift in the frequencies of light received from them and has implied much about the origins of the universe. Modern physics has been profoundly affected by observations of Doppler shifts.

For a stationary observer and a moving source, the frequency f_{obs} received by the observer can be shown to be

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right),$$

where f_{s} is the frequency of the source, v_{s} is the speed of the source along a line joining the source and observer, and v_{w} is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away from the observer, producing the appropriate shifts up and down in frequency. Note that the greater the speed of the source, the greater the effect. Similarly, for a stationary source and moving observer, the frequency received by the observer f_{obs} is given by

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right),$$

where v_{obs} is the speed of the observer along a line joining the source and observer. Here the plus sign is for motion toward the source, and the minus is for motion away from the source.

Example:

Calculate Doppler Shift: A Train Horn

Suppose a train that has a 150-Hz horn is moving at 35.0 m/s in still air on a day when the speed of sound is 340 m/s.

(a) What frequencies are observed by a stationary person at the side of the tracks as the train approaches and after it passes?

(b) What frequency is observed by the train’s engineer traveling on the train?

Strategy

To find the observed frequency in (a), $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$, must be used because the source is moving. The minus sign is used for the approaching train, and the plus sign for the receding train. In (b), there are two Doppler shifts—one for a moving source and the other for a moving observer.

Solution for (a)

(1) Enter known values into $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{s}}}\right)$.

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} – 35.0\text{ m/s}}\right)$$

(2) Calculate the frequency observed by a stationary person as the train approaches.

Equation:

$$f_{\text{obs}} = (150\text{ Hz})(1.11) = 167\text{ Hz}$$

(3) Use the same equation with the plus sign to find the frequency heard by a stationary person as the train recedes.

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} + v_{\text{s}}}\right) = (150\text{ Hz})\left(\frac{340\text{ m/s}}{340\text{ m/s} + 35.0\text{ m/s}}\right)$$

(4) Calculate the second frequency.

Equation:

$$f_{\text{obs}} = (150\text{ Hz})(0.907) = 136\text{ Hz}$$

Discussion on (a)

The numbers calculated are valid when the train is far enough away that the motion is nearly along the line joining train and observer. In both cases, the shift is significant and easily noticed. Note that the shift is 17.0 Hz for motion toward and 14.0 Hz for motion away. The shifts are not symmetric.

Solution for (b)

(1) Identify knowns:

- It seems reasonable that the engineer would receive the same frequency as emitted by the horn, because the relative velocity

between them is zero.
- Relative to the medium (air), the speeds are $v_{\text{s}} = v_{\text{obs}} = 35.0 \text{ m/s}$.
- The first Doppler shift is for the moving observer; the second is for the moving source.

(2) Use the following equation:

Equation:

$$f_{\text{obs}} = \left[f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right)\right]\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right).$$

The quantity in the square brackets is the Doppler-shifted frequency due to a moving observer. The factor on the right is the effect of the moving source.

(3) Because the train engineer is moving in the direction toward the horn, we must use the plus sign for v_{obs}; however, because the horn is also moving in the direction away from the engineer, we also use the plus sign for v_{s}. But the train is carrying both the engineer and the horn at the same velocity, so $v_{\text{s}} = v_{\text{obs}}$. As a result, everything but f_{s} cancels, yielding

Equation:

$$f_{\text{obs}}{=}f_{\text{s}}.$$

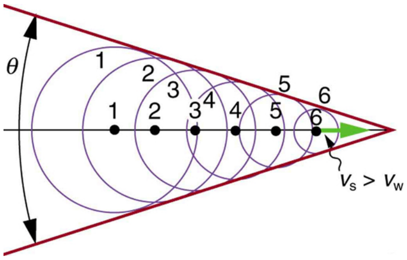
Discussion for (b)

We may expect that there is no change in frequency when source and observer move together because it fits your experience. For example, there is no Doppler shift in the frequency of conversations between driver and passenger on a motorcycle. People talking when a wind moves the air between them also observe no Doppler shift in their conversation. The crucial point is that source and observer are not moving relative to each other.

Sonic Booms to Bow Wakes

What happens to the sound produced by a moving source, such as a jet airplane, that approaches or even exceeds the speed of sound? The answer to this question applies not only to sound but to all other waves as well.

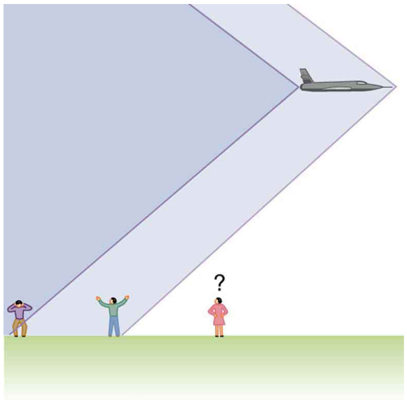
Suppose a jet airplane is coming nearly straight at you, emitting a sound of frequency f_s. The greater the plane's speed v_s, the greater the Doppler shift and the greater the value observed for f_obs. Now, as v_s approaches the speed of sound, f_obs approaches infinity, because the denominator in $f_\text{obs} = f_\text{s}\left(\frac{v_\text{w}}{v_\text{w} \pm v_\text{s}}\right)$ approaches zero. At the speed of sound, this result means that in front of the source, each successive wave is superimposed on the previous one because the source moves forward at the speed of sound. The observer gets them all at the same instant, and so the frequency is infinite. (Before airplanes exceeded the speed of sound, some people argued it would be impossible because such constructive superposition would produce pressures great enough to destroy the airplane.) If the source exceeds the speed of sound, no sound is received by the observer until the source has passed, so that the sounds from the approaching source are mixed with those from it when receding. This mixing appears messy, but something interesting happens—a sonic boom is created. (See [link].)



Sound waves from a source that moves faster than the speed of sound spread spherically from the point where they are emitted, but the source moves ahead of each.

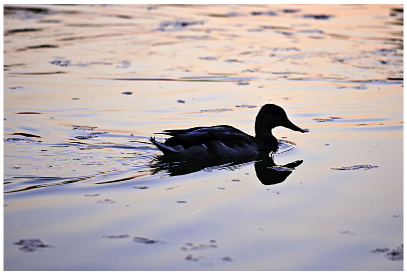
Constructive interference along the lines shown (actually a cone in three dimensions) creates a shock wave called a sonic boom. The faster the speed of the source, the smaller the angle θ.

There is constructive interference along the lines shown (a cone in three dimensions) from similar sound waves arriving there simultaneously. This superposition forms a disturbance called a **sonic boom**, a constructive interference of sound created by an object moving faster than sound. Inside the cone, the interference is mostly destructive, and so the sound intensity there is much less than on the shock wave. An aircraft creates two sonic booms, one from its nose and one from its tail. (See [link].) During television coverage of space shuttle landings, two distinct booms could often be heard. These were separated by exactly the time it would take the shuttle to pass by a point. Observers on the ground often do not see the aircraft creating the sonic boom, because it has passed by before the shock wave reaches them, as seen in [link]. If the aircraft flies close by at low altitude, pressures in the sonic boom can be destructive and break windows as well as rattle nerves. Because of how destructive sonic booms can be, supersonic flights are banned over populated areas of the United States.

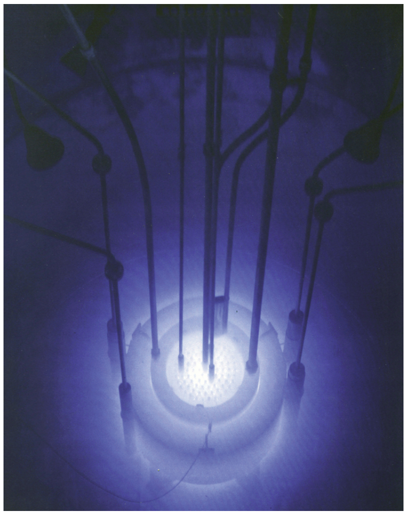


Two sonic booms, created by the nose and tail of an aircraft, are observed on the ground after the plane has passed by.

Sonic booms are one example of a broader phenomenon called bow wakes. A **bow wake**, such as the one in [link], is created when the wave source moves faster than the wave propagation speed. Water waves spread out in circles from the point where created, and the bow wake is the familiar V-shaped wake trailing the source. A more exotic bow wake is created when a subatomic particle travels through a medium faster than the speed of light travels in that medium. (In a vacuum, the maximum speed of light will be $c = 3.00 \times 10^8 \text{ m/s}$; in the medium of water, the speed of light is closer to $0.75c$. If the particle creates light in its passage, that light spreads on a cone with an angle indicative of the speed of the particle, as illustrated in [link]. Such a bow wake is called Cerenkov radiation and is commonly observed in particle physics.



Bow wake created by a duck. Constructive interference produces the rather structured wake, while there is relatively little wave action inside the wake, where interference is mostly destructive. (credit: Horia Varlan, Flickr)



The blue glow in this research reactor pool is Cerenkov radiation caused by subatomic particles traveling faster than the speed of light in water. (credit: U.S. Nuclear Regulatory Commission)

Doppler shifts and sonic booms are interesting sound phenomena that occur in all types of waves. They can be of considerable use. For example, the Doppler shift in ultrasound can be used to measure blood velocity, while police use the Doppler shift in radar (a microwave) to measure car velocities. In meteorology, the Doppler shift is used to track the motion of storm clouds; such “Doppler Radar” can give velocity and direction and rain or snow potential of imposing weather fronts. In astronomy, we can examine the light emitted from distant galaxies and determine their speed relative to ours. As galaxies move away from us, their light is shifted to a lower frequency, and so to a longer wavelength—the so-called red shift. Such information from galaxies far, far away has allowed us to estimate the age of the universe (from the Big Bang) as about 14 billion years.

Exercise:

Check Your Understanding

Problem:

Why did scientist Christian Doppler observe musicians both on a moving train and also from a stationary point not on the train?

Solution:

Doppler needed to compare the perception of sound when the observer is stationary and the sound source moves, as well as when the sound source and the observer are both in motion.

Exercise:
Check Your Understanding

Problem:

Describe a situation in your life when you might rely on the Doppler shift to help you either while driving a car or walking near traffic.

Solution:

If I am driving and I hear Doppler shift in an ambulance siren, I would be able to tell when it was getting closer and also if it has passed by. This would help me to know whether I needed to pull over and let the ambulance through.

Section Summary

- The Doppler effect is an alteration in the observed frequency of a sound due to motion of either the source or the observer.
- The actual change in frequency is called the Doppler shift.
- A sonic boom is constructive interference of sound created by an object moving faster than sound.
- A sonic boom is a type of bow wake created when any wave source moves faster than the wave propagation speed.
- For a stationary observer and a moving source, the observed frequency f_{obs} is:
  Equation:

  $$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right),$$

  where f_{s} is the frequency of the source, v_{s} is the speed of the source, and v_{w} is the speed of sound. The minus sign is used for motion toward the observer and the plus sign for motion away.
- For a stationary source and moving observer, the observed frequency is:

Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right),$$

where v_{obs} is the speed of the observer.

Conceptual Questions

Exercise:

Problem: Is the Doppler shift real or just a sensory illusion?

Exercise:

Problem:

Due to efficiency considerations related to its bow wake, the supersonic transport aircraft must maintain a cruising speed that is a constant ratio to the speed of sound (a constant Mach number). If the aircraft flies from warm air into colder air, should it increase or decrease its speed? Explain your answer.

Exercise:

Problem:

When you hear a sonic boom, you often cannot see the plane that made it. Why is that?

Problems & Exercises

Exercise:

Problem:

(a) What frequency is received by a person watching an oncoming ambulance moving at 110 km/h and emitting a steady 800-Hz sound from its siren? The speed of sound on this day is 345 m/s. (b) What frequency does she receive after the ambulance has passed?

Solution:

(a) 878 Hz

(b) 735 Hz

Exercise:

Problem:

(a) At an air show a jet flies directly toward the stands at a speed of 1200 km/h, emitting a frequency of 3500 Hz, on a day when the speed of sound is 342 m/s. What frequency is received by the observers? (b) What frequency do they receive as the plane flies directly away from them?

Exercise:

Problem:

What frequency is received by a mouse just before being dispatched by a hawk flying at it at 25.0 m/s and emitting a screech of frequency 3500 Hz? Take the speed of sound to be 331 m/s.

Solution:
Equation:

$$3.79 \times 10^3 \text{ Hz}$$

Exercise:

Problem:

A spectator at a parade receives an 888-Hz tone from an oncoming trumpeter who is playing an 880-Hz note. At what speed is the musician approaching if the speed of sound is 338 m/s?

Exercise:

Problem:

A commuter train blows its 200-Hz horn as it approaches a crossing. The speed of sound is 335 m/s. (a) An observer waiting at the crossing receives a frequency of 208 Hz. What is the speed of the train? (b) What frequency does the observer receive as the train moves away?

Solution:

(a) 12.9 m/s

(b) 193 Hz

Exercise:

Problem:

Can you perceive the shift in frequency produced when you pull a tuning fork toward you at 10.0 m/s on a day when the speed of sound is 344 m/s? To answer this question, calculate the factor by which the frequency shifts and see if it is greater than 0.300%.

Exercise:

Problem:

Two eagles fly directly toward one another, the first at 15.0 m/s and the second at 20.0 m/s. Both screech, the first one emitting a frequency of 3200 Hz and the second one emitting a frequency of 3800 Hz. What frequencies do they receive if the speed of sound is 330 m/s?

Solution:

First eagle hears $4.23 \times 10^3 \text{ Hz}$

Second eagle hears $3.56 \times 10^3 \text{ Hz}$

Exercise:

Problem:

What is the minimum speed at which a source must travel toward you for you to be able to hear that its frequency is Doppler shifted? That is, what speed produces a shift of 0.300% on a day when the speed of sound is 331 m/s?

Glossary

Doppler effect
: an alteration in the observed frequency of a sound due to motion of either the source or the observer

Doppler shift
: the actual change in frequency due to relative motion of source and observer

sonic boom
: a constructive interference of sound created by an object moving faster than sound

bow wake
: V-shaped disturbance created when the wave source moves faster than the wave propagation speed

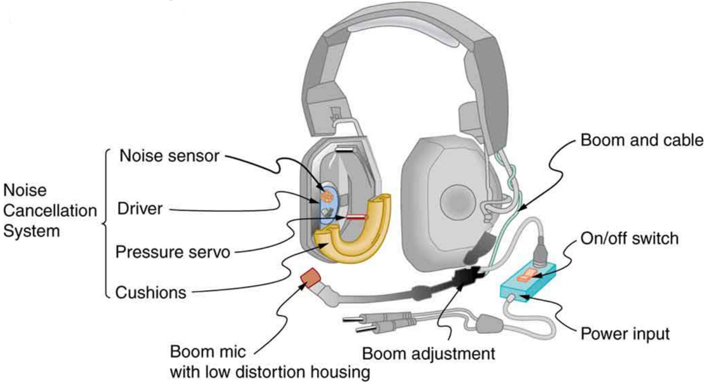
Sound Interference and Resonance: Standing Waves in Air Columns

- Define antinode, node, fundamental, overtones, and harmonics.
- Identify instances of sound interference in everyday situations.
- Describe how sound interference occurring inside open and closed tubes changes the characteristics of the sound, and how this applies to sounds produced by musical instruments.
- Calculate the length of a tube using sound wave measurements.

Some types of headphones use the phenomena of constructive and destructive interference to cancel out outside noises. (credit: JVC America, Flickr)

Interference is the hallmark of waves, all of which exhibit constructive and destructive interference exactly analogous to that seen for water waves. In fact, one way to prove something “is a wave” is to observe interference effects. So, sound being a wave, we expect it to exhibit interference; we have already mentioned a few such effects, such as the beats from two similar notes played simultaneously.

[link] shows a clever use of sound interference to cancel noise. Larger-scale applications of active noise reduction by destructive interference are contemplated for entire passenger compartments in commercial aircraft. To obtain destructive interference, a fast electronic analysis is performed, and a second sound is introduced with its maxima and minima exactly reversed from the incoming noise. Sound waves in fluids are pressure waves and consistent with Pascal’s principle; pressures from two different sources add and subtract like simple numbers; that is, positive and negative gauge pressures add to a much smaller pressure, producing a lower-intensity sound. Although completely destructive interference is possible only under the simplest conditions, it is possible to reduce noise levels by 30 dB or more using this technique.

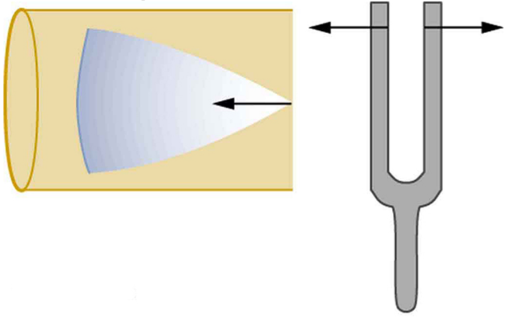


Headphones designed to cancel noise with destructive interference create a sound wave exactly opposite to the incoming sound. These headphones can be more effective than the simple passive attenuation used in most ear protection. Such headphones were used on the record-setting, around the world nonstop flight of the Voyager aircraft to protect the pilots’ hearing from engine noise.

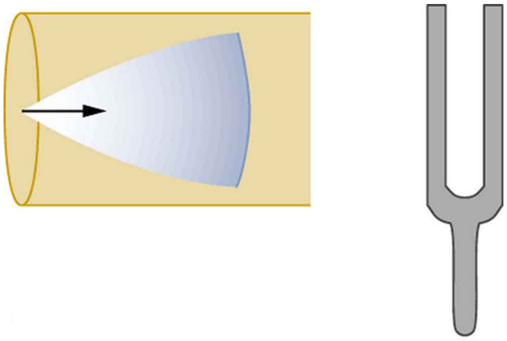
Where else can we observe sound interference? All sound resonances, such as in musical instruments, are due to constructive and destructive interference. Only the resonant frequencies interfere constructively to form standing waves, while others interfere destructively and are absent. From the toot made by blowing over a bottle, to the characteristic flavor of a violin’s sounding box, to the recognizability of a great singer’s voice, resonance and standing waves play a vital role.

Note:
Interference
Interference is such a fundamental aspect of waves that observing interference is proof that something is a wave. The wave nature of light was established by experiments showing interference. Similarly, when electrons scattered from crystals exhibited interference, their wave nature was confirmed to be exactly as predicted by symmetry with certain wave characteristics of light.

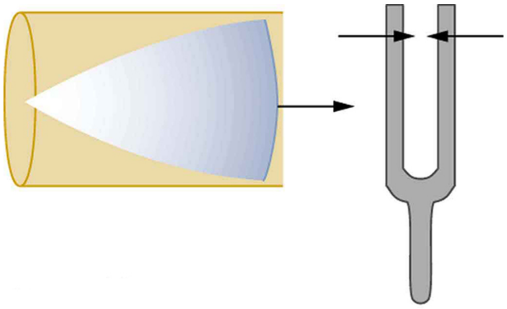
Suppose we hold a tuning fork near the end of a tube that is closed at the other end, as shown in [link], [link], [link], and [link]. If the tuning fork has just the right frequency, the air column in the tube resonates loudly, but at most frequencies it vibrates very little. This observation just means that the air column has only certain natural frequencies. The figures show how a resonance at the lowest of these natural frequencies is formed. A disturbance travels down the tube at the speed of sound and bounces off the closed end. If the tube is just the right length, the reflected sound arrives back at the tuning fork exactly half a cycle later, and it interferes constructively with the continuing sound produced by the tuning fork. The incoming and reflected sounds form a standing wave in the tube as shown.



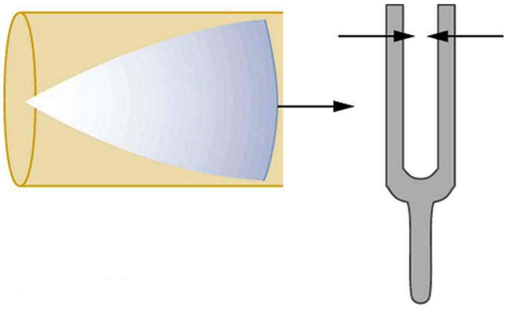
Resonance of air in a tube closed at one end, caused by a tuning fork. A disturbance moves down the tube.



Resonance of air in a tube closed at one end, caused by a tuning fork. The disturbance reflects from the closed end of the tube.

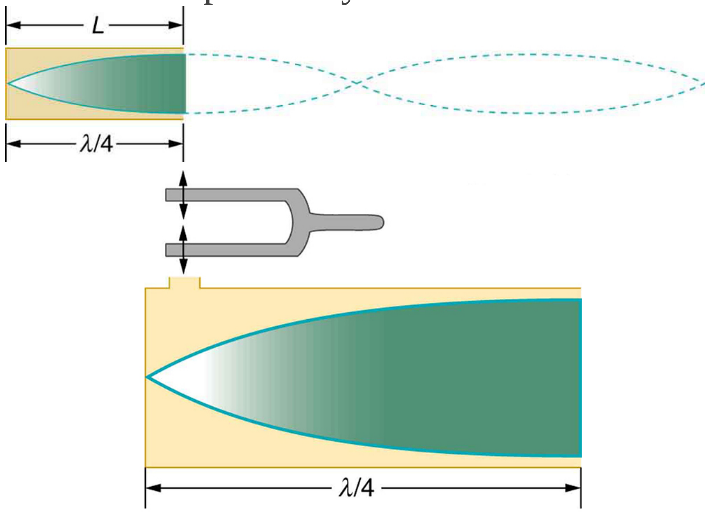


Resonance of air in a tube closed at one end, caused by a tuning fork. If the length of the tube L is just right, the disturbance gets back to the tuning fork half a cycle later and interferes constructively with the continuing sound from the tuning fork. This interference forms a standing wave, and the air column resonates.



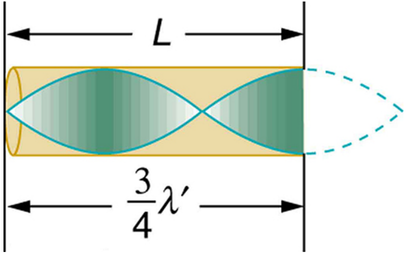
Resonance of air in a tube closed at one end, caused by a tuning fork. A graph of air displacement along the length of the tube shows none at the closed end, where the motion is constrained, and a maximum at the open end. This standing wave has one-fourth of its wavelength in the tube, so that $\lambda = 4L$.

The standing wave formed in the tube has its maximum air displacement (an **antinode**) at the open end, where motion is unconstrained, and no displacement (a **node**) at the closed end, where air movement is halted. The distance from a node to an antinode is one-fourth of a wavelength, and this equals the length of the tube; thus, $\lambda = 4L$. This same resonance can be produced by a vibration introduced at or near the closed end of the tube, as shown in [link]. It is best to consider this a natural vibration of the air column independently of how it is induced.

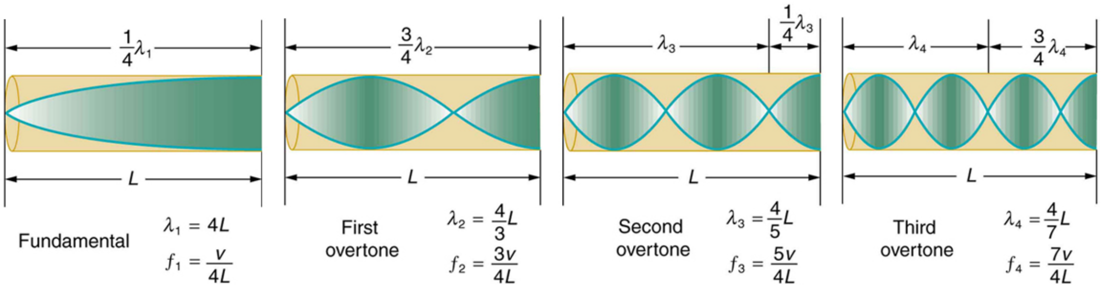


The same standing wave is created in the tube by a vibration introduced near its closed end.

Given that maximum air displacements are possible at the open end and none at the closed end, there are other, shorter wavelengths that can resonate in the tube, such as the one shown in [link]. Here the standing wave has three-fourths of its wavelength in the tube, or $L=(3/4)\lambda\prime$, so that $\lambda\prime=4L/3$. Continuing this process reveals a whole series of shorter-wavelength and higher-frequency sounds that resonate in the tube. We use specific terms for the resonances in any system. The lowest resonant frequency is called the **fundamental**, while all higher resonant frequencies are called **overtones**. All resonant frequencies are integral multiples of the fundamental, and they are collectively called **harmonics**. The fundamental is the first harmonic, the first overtone is the second harmonic, and so on. [link] shows the fundamental and the first three overtones (the first four harmonics) in a tube closed at one end.

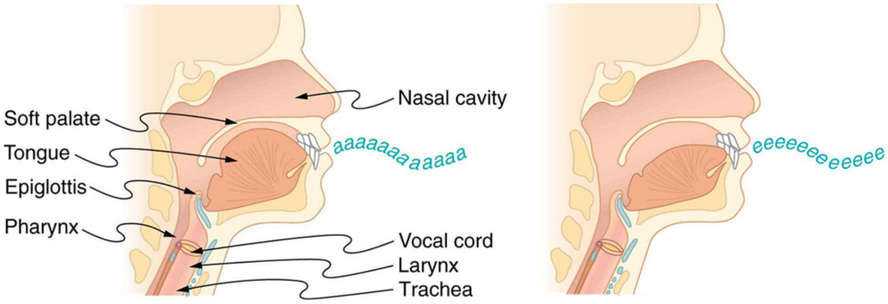


Another resonance for a tube closed at one end. This has maximum air displacements at the open end, and none at the closed end. The wavelength is shorter, with three-fourths $\lambda\prime$ equaling the length of the tube, so that $\lambda\prime=4L/3$. This higher-frequency vibration is the first overtone.



The fundamental and three lowest overtones for a tube closed at one end. All have maximum air displacements at the open end and none at the closed end.

The fundamental and overtones can be present simultaneously in a variety of combinations. For example, middle C on a trumpet has a sound distinctively different from middle C on a clarinet, both instruments being modified versions of a tube closed at one end. The fundamental frequency is the same (and usually the most intense), but the overtones and their mix of intensities are different and subject to shading by the musician. This mix is what gives various musical instruments (and human voices) their distinctive characteristics, whether they have air columns, strings, sounding boxes, or drumheads. In fact, much of our speech is determined by shaping the cavity formed by the throat and mouth and positioning the tongue to adjust the fundamental and combination of overtones. Simple resonant cavities can be made to resonate with the sound of the vowels, for example. (See [link].) In boys, at puberty, the larynx grows and the shape of the resonant cavity changes giving rise to the difference in predominant frequencies in speech between men and women.



The throat and mouth form an air column closed at one end that resonates in response to vibrations in the voice box. The spectrum of overtones and their intensities vary with mouth shaping and tongue position to form different sounds. The voice box can be replaced with a mechanical vibrator, and understandable speech is still possible. Variations in basic shapes make different voices recognizable.

Now let us look for a pattern in the resonant frequencies for a simple tube that is closed at one end. The fundamental has $\lambda = 4L$, and frequency is related to wavelength and the speed of sound as given by:

Equation:

$$v_{\text{w}} = f\lambda.$$

Solving for f in this equation gives

Equation:

$$f = \frac{v_{\text{w}}}{\lambda} = \frac{v_{\text{w}}}{4L},$$

where v_{w} is the speed of sound in air. Similarly, the first overtone has $\lambda\prime = 4L/3$ (see [link]), so that

Equation:

$$f\prime = 3\frac{v_{\text{w}}}{4L} = 3f.$$

Because $f\prime = 3f$, we call the first overtone the third harmonic. Continuing this process, we see a pattern that can be generalized in a single expression. The resonant frequencies of a tube closed at one end are

Equation:

$$f_n = n\frac{v_\text{w}}{4L}, n = 1,3,5,$$

where f_1 is the fundamental, f_3 is the first overtone, and so on. It is interesting that the resonant frequencies depend on the speed of sound and, hence, on temperature. This dependence poses a noticeable problem for organs in old unheated cathedrals, and it is also the reason why musicians commonly bring their wind instruments to room temperature before playing them.

Example:
Find the Length of a Tube with a 128 Hz Fundamental

(a) What length should a tube closed at one end have on a day when the air temperature, is 22.0°C, if its fundamental frequency is to be 128 Hz (C below middle C)?

(b) What is the frequency of its fourth overtone?

Strategy

The length L can be found from the relationship in $f_n = n\frac{v_\text{w}}{4L}$, but we will first need to find the speed of sound v_w.

Solution for (a)

(1) Identify knowns:

- the fundamental frequency is 128 Hz
- the air temperature is 22.0°C

(2) Use $f_n = n\frac{v_\text{w}}{4L}$ to find the fundamental frequency ($n = 1$).

Equation:

$$f_1 = \frac{v_\text{w}}{4L}$$

(3) Solve this equation for length.

Equation:

$$L = \frac{v_\text{w}}{4f_1}$$

(4) Find the speed of sound using $v_{\mathrm{w}}=(331\ \mathrm{m/s})\sqrt{\frac{T}{273\ \mathrm{K}}}$.

Equation:

$$v_{\mathrm{w}}=(331\ \mathrm{m/s})\sqrt{\frac{295\ \mathrm{K}}{273\ \mathrm{K}}}=344\ \mathrm{m/s}$$

(5) Enter the values of the speed of sound and frequency into the expression for L.

Equation:

$$L=\frac{v_{\mathrm{w}}}{4f_1}=\frac{344\ \mathrm{m/s}}{4(128\ \mathrm{Hz})}=0.672\ \mathrm{m}$$

Discussion on (a)

Many wind instruments are modified tubes that have finger holes, valves, and other devices for changing the length of the resonating air column and hence, the frequency of the note played. Horns producing very low frequencies, such as tubas, require tubes so long that they are coiled into loops.

Solution for (b)

(1) Identify knowns:

- the first overtone has $n=3$
- the second overtone has $n=5$
- the third overtone has $n=7$
- the fourth overtone has $n=9$

(2) Enter the value for the fourth overtone into $f_n=n\frac{v_{\mathrm{w}}}{4L}$.

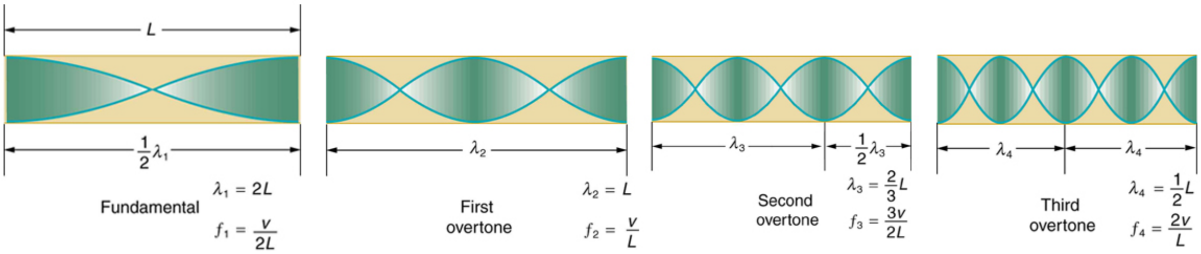
Equation:

$$f_9=9\frac{v_{\mathrm{w}}}{4L}=9f_1=1.15\ \mathrm{kHz}$$

Discussion on (b)

Whether this overtone occurs in a simple tube or a musical instrument depends on how it is stimulated to vibrate and the details of its shape. The trombone, for example, does not produce its fundamental frequency and only makes overtones.

Another type of tube is one that is *open* at both ends. Examples are some organ pipes, flutes, and oboes. The resonances of tubes open at both ends can be analyzed in a very similar fashion to those for tubes closed at one end. The air columns in tubes open at both ends have maximum air displacements at both ends, as illustrated in [link]. Standing waves form as shown.



The resonant frequencies of a tube open at both ends are shown, including the fundamental and the first three overtones. In all cases the maximum air displacements occur at both ends of the tube, giving it different natural frequencies than a tube closed at one end.

Based on the fact that a tube open at both ends has maximum air displacements at both ends, and using [link] as a guide, we can see that the resonant frequencies of a tube open at both ends are:

Equation:

$$f_n = n\frac{v_{\text{w}}}{2L},\ n = 1, 2, 3...,$$

where f_1 is the fundamental, f_2 is the first overtone, f_3 is the second overtone, and so on. Note that a tube open at both ends has a fundamental frequency twice what it would have if closed at one end. It also has a different spectrum of overtones than a tube closed at one end. So if you had two tubes with the same fundamental frequency but one was open at both ends and the other was closed at one end, they would sound different when played because they have different overtones. Middle C, for example, would sound richer played on an open tube, because it has even multiples of the fundamental as well as odd. A closed tube has only odd multiples.

Note:

Real-World Applications: Resonance in Everyday Systems

Resonance occurs in many different systems, including strings, air columns, and atoms. Resonance is the driven or forced oscillation of a system at its natural frequency. At resonance, energy is transferred rapidly to the oscillating system, and the amplitude of its oscillations grows until the system can no longer be described by Hooke’s law. An example of this is the distorted sound intentionally produced in certain types of rock music.

Wind instruments use resonance in air columns to amplify tones made by lips or vibrating reeds. Other instruments also use air resonance in clever ways to amplify sound. [link] shows a violin and a guitar, both of which have sounding boxes but with different shapes, resulting in different overtone structures. The vibrating string creates a sound that resonates in the sounding box, greatly amplifying the sound and creating overtones that give the instrument its characteristic flavor. The more complex the shape of the sounding box, the greater its ability to resonate over a wide range of frequencies. The marimba, like the one shown in [link] uses pots or gourds below the wooden slats to amplify their tones. The resonance of the pot can be adjusted by adding water.

String instruments such as violins and guitars use resonance in their sounding boxes to amplify and enrich the sound created by their vibrating strings. The bridge and supports couple the string vibrations to the sounding boxes and air within. (credits: guitar, Feliciano Guimares, Fotopedia; violin, Steve Snodgrass, Flickr)

Resonance has been used in musical instruments since prehistoric times. This marimba uses gourds as resonance chambers to amplify its sound. (credit: APC Events, Flickr)

We have emphasized sound applications in our discussions of resonance and standing waves, but these ideas apply to any system that has wave characteristics. Vibrating strings, for example, are actually resonating and have fundamentals and overtones similar to those for air columns. More subtle are the resonances in atoms due to the wave character of their electrons. Their orbitals can be viewed as standing waves, which have a fundamental (ground state) and overtones (excited states). It is fascinating that wave characteristics apply to such a wide range of physical systems.

Exercise:

Check Your Understanding

Problem:

Describe how noise-canceling headphones differ from standard headphones used to block outside sounds.

Solution:

Regular headphones only block sound waves with a physical barrier. Noise-canceling headphones use destructive interference to reduce the loudness of outside sounds.

Exercise:
Check Your Understanding

Problem:

How is it possible to use a standing wave's node and antinode to determine the length of a closed-end tube?

Solution:

When the tube resonates at its natural frequency, the wave's node is located at the closed end of the tube, and the antinode is located at the open end. The length of the tube is equal to one-fourth of the wavelength of this wave. Thus, if we know the wavelength of the wave, we can determine the length of the tube.

Note:
PhET Explorations: Sound
This simulation lets you see sound waves. Adjust the frequency or volume and you can see and hear how the wave changes. Move the listener around and hear what she hears.
https://archive.cnx.org/specials/c4d3b96e-41f3-11e5-ab7b-47e22dffc18e/sound/#sim-single-source

Section Summary

- Sound interference and resonance have the same properties as defined for all waves.
- In air columns, the lowest-frequency resonance is called the fundamental, whereas all higher resonant frequencies are called overtones. Collectively, they are called harmonics.

- The resonant frequencies of a tube closed at one end are:
  Equation:

  $$f_n = n\frac{v_\text{w}}{4L}, n = 1, 3, 5...,$$

  f_1 is the fundamental and L is the length of the tube.
- The resonant frequencies of a tube open at both ends are:
  Equation:

  $$f_n = n\frac{v_\text{w}}{2L}, n = 1, 2, 3...$$

Conceptual Questions

Exercise:

Problem:

How does an unamplified guitar produce sounds so much more intense than those of a plucked string held taut by a simple stick?

Exercise:

Problem:

You are given two wind instruments of identical length. One is open at both ends, whereas the other is closed at one end. Which is able to produce the lowest frequency?

Exercise:

Problem:

What is the difference between an overtone and a harmonic? Are all harmonics overtones? Are all overtones harmonics?

Problems & Exercises

Exercise:

Problem:

A “showy” custom-built car has two brass horns that are supposed to produce the same frequency but actually emit 263.8 and 264.5 Hz. What beat frequency is produced?

Solution:

0.7 Hz

Exercise:

Problem:

What beat frequencies will be present: (a) If the musical notes A and C are played together (frequencies of 220 and 264 Hz)? (b) If D and F are played together (frequencies of 297 and 352 Hz)? (c) If all four are played together?

Exercise:

Problem:

What beat frequencies result if a piano hammer hits three strings that emit frequencies of 127.8, 128.1, and 128.3 Hz?

Solution:

0.3 Hz, 0.2 Hz, 0.5 Hz

Exercise:

Problem:

A piano tuner hears a beat every 2.00 s when listening to a 264.0-Hz tuning fork and a single piano string. What are the two possible frequencies of the string?

Exercise:

Problem:

(a) What is the fundamental frequency of a 0.672-m-long tube, open at both ends, on a day when the speed of sound is 344 m/s? (b) What is the frequency of its second harmonic?

Solution:

(a) 256 Hz

(b) 512 Hz

Exercise:

Problem:

If a wind instrument, such as a tuba, has a fundamental frequency of 32.0 Hz, what are its first three overtones? It is closed at one end. (The overtones of a real tuba are more complex than this example, because it is a tapered tube.)

Exercise:

Problem:

What are the first three overtones of a bassoon that has a fundamental frequency of 90.0 Hz? It is open at both ends. (The overtones of a real bassoon are more complex than this example, because its double reed makes it act more like a tube closed at one end.)

Solution:

180 Hz, 270 Hz, 360 Hz

Exercise:

Problem:

How long must a flute be in order to have a fundamental frequency of 262 Hz (this frequency corresponds to middle C on the evenly tempered chromatic scale) on a day when air temperature is 20.0°C? It is open at both ends.

Exercise:

Problem:

What length should an oboe have to produce a fundamental frequency of 110 Hz on a day when the speed of sound is 343 m/s? It is open at both ends.

Solution:

1.56 m

Exercise:

Problem:

What is the length of a tube that has a fundamental frequency of 176 Hz and a first overtone of 352 Hz if the speed of sound is 343 m/s?

Exercise:

Problem:

(a) Find the length of an organ pipe closed at one end that produces a fundamental frequency of 256 Hz when air temperature is 18.0°C. (b) What is its fundamental frequency at 25.0°C?

Solution:

(a) 0.334 m

(b) 259 Hz

Exercise:

Problem:

By what fraction will the frequencies produced by a wind instrument change when air temperature goes from 10.0°C to 30.0°C? That is, find the ratio of the frequencies at those temperatures.

Exercise:

Problem:

The ear canal resonates like a tube closed at one end. (See [link].) If ear canals range in length from 1.80 to 2.60 cm in an average population, what is the range of fundamental resonant frequencies? Take air temperature to be 37.0°C, which is the same as body temperature. How does this result correlate with the intensity versus frequency graph ([link] of the human ear?

Solution:

3.39 to 4.90 kHz

Exercise:

Problem:

Calculate the first overtone in an ear canal, which resonates like a 2.40-cm-long tube closed at one end, by taking air temperature to be 37.0°C. Is the ear particularly sensitive to such a frequency? (The resonances of the ear canal are complicated by its nonuniform shape, which we shall ignore.)

Exercise:

Problem:

A crude approximation of voice production is to consider the breathing passages and mouth to be a resonating tube closed at one end. (See [link].) (a) What is the fundamental frequency if the tube is 0.240-m long, by taking air temperature to be 37.0°C? (b) What would this frequency become if the person replaced the air with helium? Assume the same temperature dependence for helium as for air.

Solution:

(a) 367 Hz

(b) 1.07 kHz

Exercise:

Problem:

(a) Students in a physics lab are asked to find the length of an air column in a tube closed at one end that has a fundamental frequency of 256 Hz. They hold the tube vertically and fill it with water to the top, then lower the water while a 256-Hz tuning fork is rung and listen for the first resonance. What is the air temperature if the resonance occurs for a length of 0.336 m? (b) At what length will they observe the second resonance (first overtone)?

Exercise:

Problem:

What frequencies will a 1.80-m-long tube produce in the audible range at 20.0°C if: (a) The tube is closed at one end? (b) It is open at both ends?

Solution:

(a) $f_n = n(47.6\text{ Hz}),\ n = 1, 3, 5,..., 419$

(b) $f_n = n(95.3\text{ Hz}),\ n = 1, 2, 3,..., 210$

Glossary

antinode
: point of maximum displacement

node

point of zero displacement

fundamental
the lowest-frequency resonance

overtones
all resonant frequencies higher than the fundamental

harmonics
the term used to refer collectively to the fundamental and its overtones

Hearing

- Define hearing, pitch, loudness, timbre, note, tone, phon, ultrasound, and infrasound.
- Compare loudness to frequency and intensity of a sound.
- Identify structures of the inner ear and explain how they relate to sound perception.

Hearing allows this vocalist, his band, and his fans to enjoy music. (credit: West Point Public Affairs, Flickr)

The human ear has a tremendous range and sensitivity. It can give us a wealth of simple information—such as pitch, loudness, and direction. And from its input we can detect musical quality and nuances of voiced emotion. How is our hearing related to the physical qualities of sound, and how does the hearing mechanism work?

Hearing is the perception of sound. (Perception is commonly defined to be awareness through the senses, a typically circular definition of higher-level processes in living organisms.) Normal human hearing encompasses frequencies from 20 to 20,000 Hz, an impressive range. Sounds below 20 Hz are called **infrasound**, whereas those above 20,000 Hz are **ultrasound**. Neither is perceived by the ear, although infrasound can sometimes be felt as vibrations. When we do hear low-frequency vibrations, such as the sounds of a diving board, we hear the individual vibrations only because there are higher-frequency sounds in each. Other animals have hearing ranges different from that of humans. Dogs can hear sounds as high as 30,000 Hz, whereas bats and dolphins can hear up to 100,000-Hz sounds. You may have noticed that dogs respond to the sound of a dog whistle which produces sound out of the range of human hearing. Elephants are known to respond to frequencies below 20 Hz.

The perception of frequency is called **pitch**. Most of us have excellent relative pitch, which means that we can tell whether one sound has a different frequency from another. Typically, we can discriminate between two sounds if their frequencies differ by 0.3% or more. For example, 500.0 and 501.5 Hz are noticeably different. Pitch perception is directly related to frequency and is not greatly affected by other physical quantities such as intensity. Musical **notes** are particular sounds that can be produced by most instruments and in Western music have particular names. Combinations of notes constitute music. Some people can identify musical notes, such as A-sharp, C, or E-flat, just by listening to them. This uncommon ability is called perfect pitch.

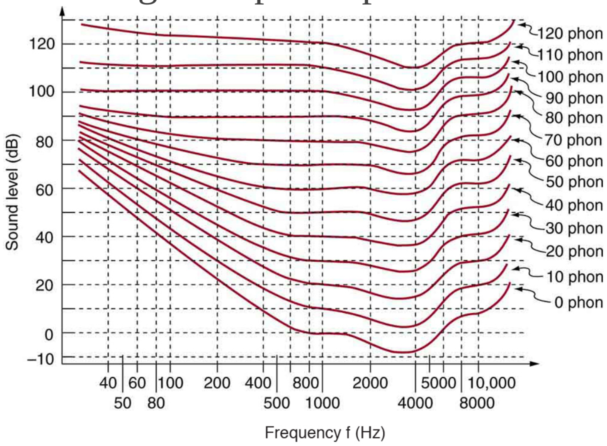
The ear is remarkably sensitive to low-intensity sounds. The lowest audible intensity or threshold is about $10^{-12}\ \text{W/m}^2$ or 0 dB. Sounds as much as 10^{12} more intense can be briefly tolerated. Very few measuring devices are capable of observations over a range of a trillion. The perception of intensity is called **loudness**. At a given frequency, it is possible to discern differences of about 1 dB, and a change of 3 dB is easily noticed. But loudness is not related to intensity alone. Frequency has a major effect on how loud a sound seems. The ear has its maximum sensitivity to frequencies in the range of 2000 to 5000 Hz, so that sounds in this range are perceived as being louder than, say, those at 500 or 10,000 Hz, even when they all have the same intensity. Sounds near the high- and low-frequency extremes of the hearing range seem even less loud, because the ear is even less sensitive at those frequencies. [link] gives the dependence of certain human hearing perceptions on physical quantities.

Perception	Physical quantity
Pitch	Frequency
Loudness	Intensity and Frequency
Timbre	Number and relative intensity of multiple frequencies. Subtle craftsmanship leads to non-linear effects and more detail.
Note	Basic unit of music with specific names, combined to generate tunes
Tone	Number and relative intensity of multiple frequencies.

Sound Perceptions

When a violin plays middle C, there is no mistaking it for a piano playing the same note. The reason is that each instrument produces a distinctive set of frequencies and intensities. We call our perception of these combinations of frequencies and intensities **tone** quality, or more commonly the **timbre** of the sound. It is more difficult to correlate timbre perception to physical quantities than it is for loudness or pitch perception. Timbre is more subjective. Terms such as dull, brilliant, warm, cold, pure, and rich are employed to describe the timbre of a sound. So the consideration of timbre takes us into the realm of perceptual psychology, where higher-level processes in the brain are dominant. This is true for other perceptions of sound, such as music and noise. We shall not delve further into them; rather, we will concentrate on the question of loudness perception.

A unit called a **phon** is used to express loudness numerically. Phons differ from decibels because the phon is a unit of loudness perception, whereas the decibel is a unit of physical intensity. [link] shows the relationship of loudness to intensity (or intensity level) and frequency for persons with normal hearing. The curved lines are equal-loudness curves. Each curve is labeled with its loudness in phons. Any sound along a given curve will be perceived as equally loud by the average person. The curves were determined by having large numbers of people compare the loudness of sounds at different frequencies and sound intensity levels. At a frequency of 1000 Hz, phons are taken to be numerically equal to decibels. The following example helps illustrate how to use the graph:



The relationship of loudness in phons to intensity level (in decibels) and intensity (in watts per meter squared) for persons with normal hearing. The curved lines are equal-loudness curves—all sounds on a given curve are perceived as equally loud. Phons and decibels are defined to be the same at 1000 Hz.

Example:
Measuring Loudness: Loudness Versus Intensity Level and Frequency
(a) What is the loudness in phons of a 100-Hz sound that has an intensity level of 80 dB? (b) What is the intensity level in decibels of a 4000-Hz

sound having a loudness of 70 phons? (c) At what intensity level will an 8000-Hz sound have the same loudness as a 200-Hz sound at 60 dB?

Strategy for (a)

The graph in [link] should be referenced in order to solve this example. To find the loudness of a given sound, you must know its frequency and intensity level and locate that point on the square grid, then interpolate between loudness curves to get the loudness in phons.

Solution for (a)

(1) Identify knowns:

- The square grid of the graph relating phons and decibels is a plot of intensity level versus frequency—both physical quantities.
- 100 Hz at 80 dB lies halfway between the curves marked 70 and 80 phons.

(2) Find the loudness: 75 phons.

Strategy for (b)

The graph in [link] should be referenced in order to solve this example. To find the intensity level of a sound, you must have its frequency and loudness. Once that point is located, the intensity level can be determined from the vertical axis.

Solution for (b)

(1) Identify knowns:

- Values are given to be 4000 Hz at 70 phons.

(2) Follow the 70-phon curve until it reaches 4000 Hz. At that point, it is below the 70 dB line at about 67 dB.

(3) Find the intensity level:

67 dB

Strategy for (c)

The graph in [link] should be referenced in order to solve this example.

Solution for (c)

(1) Locate the point for a 200 Hz and 60 dB sound.

(2) Find the loudness: This point lies just slightly above the 50-phon curve, and so its loudness is 51 phons.

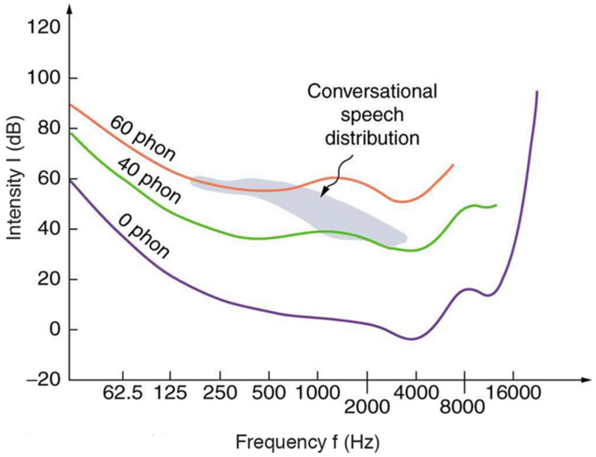
(3) Look for the 51-phon level is at 8000 Hz: 63 dB.

Discussion

These answers, like all information extracted from [link], have uncertainties of several phons or several decibels, partly due to difficulties in interpolation, but mostly related to uncertainties in the equal-loudness curves.

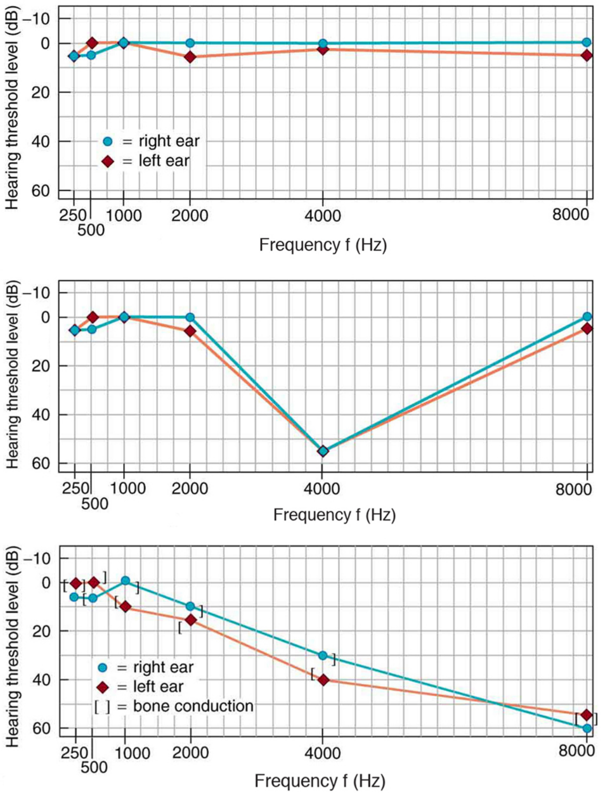
Further examination of the graph in [link] reveals some interesting facts about human hearing. First, sounds below the 0-phon curve are not perceived by most people. So, for example, a 60 Hz sound at 40 dB is inaudible. The 0-phon curve represents the threshold of normal hearing. We can hear some sounds at intensity levels below 0 dB. For example, a 3-dB, 5000-Hz sound is audible, because it lies above the 0-phon curve. The loudness curves all have dips in them between about 2000 and 5000 Hz. These dips mean the ear is most sensitive to frequencies in that range. For example, a 15-dB sound at 4000 Hz has a loudness of 20 phons, the same as a 20-dB sound at 1000 Hz. The curves rise at both extremes of the frequency range, indicating that a greater-intensity level sound is needed at those frequencies to be perceived to be as loud as at middle frequencies. For example, a sound at 10,000 Hz must have an intensity level of 30 dB to seem as loud as a 20 dB sound at 1000 Hz. Sounds above 120 phons are painful as well as damaging.

We do not often utilize our full range of hearing. This is particularly true for frequencies above 8000 Hz, which are rare in the environment and are unnecessary for understanding conversation or appreciating music. In fact, people who have lost the ability to hear such high frequencies are usually unaware of their loss until tested. The shaded region in [link] is the frequency and intensity region where most conversational sounds fall. The curved lines indicate what effect hearing losses of 40 and 60 phons will have. A 40-phon hearing loss at all frequencies still allows a person to understand conversation, although it will seem very quiet. A person with a 60-phon loss at all frequencies will hear only the lowest frequencies and will not be able to understand speech unless it is much louder than normal. Even so, speech may seem indistinct, because higher frequencies are not as well perceived. The conversational speech region also has a gender component, in that female voices are usually characterized by higher frequencies. So the person with a 60-phon hearing impediment might have difficulty understanding the normal conversation of a woman.



The shaded region represents frequencies and intensity levels found in normal conversational speech. The 0-phon line represents the normal hearing threshold, while those at 40 and 60 represent thresholds for people with 40- and 60-phon hearing losses, respectively.

Hearing tests are performed over a range of frequencies, usually from 250 to 8000 Hz, and can be displayed graphically in an audiogram like that in [link]. The hearing threshold is measured in dB *relative to the normal threshold*, so that normal hearing registers as 0 dB at all frequencies. Hearing loss caused by noise typically shows a dip near the 4000 Hz frequency, irrespective of the frequency that caused the loss and often affects both ears. The most common form of hearing loss comes with age and is called *presbycusis*—literally elder ear. Such loss is increasingly severe at higher frequencies, and interferes with music appreciation and speech recognition.

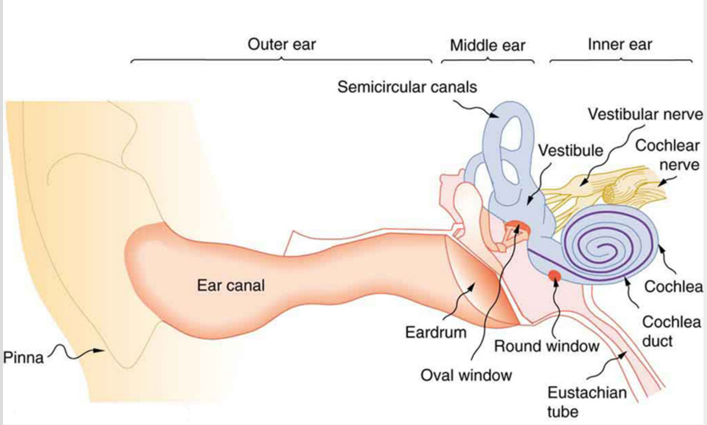


Audiograms showing the threshold in intensity level versus frequency for three different individuals. Intensity level is measured relative to the normal threshold. The top left graph is that of a person with normal hearing. The graph to its right has a dip at 4000 Hz and is that of a child who suffered hearing loss due to a cap gun. The third graph is typical of presbycusis, the progressive loss of higher frequency hearing with age. Tests performed by bone conduction (brackets) can distinguish nerve damage from middle ear damage.

Note:

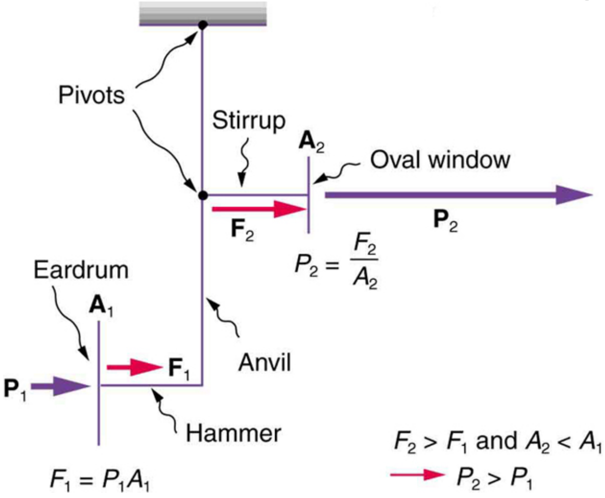
The Hearing Mechanism

The hearing mechanism involves some interesting physics. The sound wave that impinges upon our ear is a pressure wave. The ear is a transducer that converts sound waves into electrical nerve impulses in a manner much more sophisticated than, but analogous to, a microphone. [link] shows the gross anatomy of the ear with its division into three parts: the outer ear or ear canal; the middle ear, which runs from the eardrum to the cochlea; and the inner ear, which is the cochlea itself. The body part normally referred to as the ear is technically called the pinna.



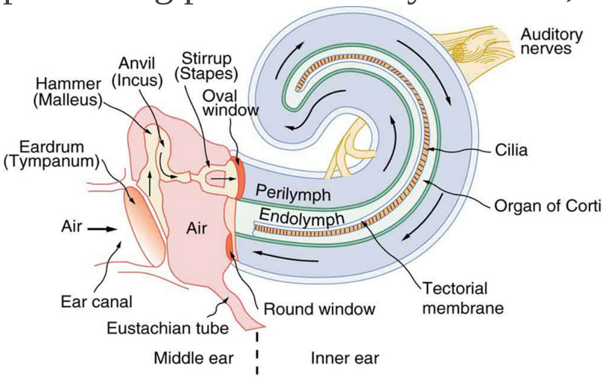
The illustration shows the gross anatomy of the human ear.

The outer ear, or ear canal, carries sound to the recessed protected eardrum. The air column in the ear canal resonates and is partially responsible for the sensitivity of the ear to sounds in the 2000 to 5000 Hz range. The middle ear converts sound into mechanical vibrations and applies these vibrations to the cochlea. The lever system of the middle ear takes the force exerted on the eardrum by sound pressure variations, amplifies it and transmits it to the inner ear via the oval window, creating pressure waves in the cochlea approximately 40 times greater than those impinging on the eardrum. (See [link].) Two muscles in the middle ear (not shown) protect the inner ear from very intense sounds. They react to intense sound in a few milliseconds and reduce the force transmitted to the cochlea. This protective reaction can also be triggered by your own voice, so that humming while shooting a gun, for example, can reduce noise damage.



This schematic shows the middle ear’s system for converting sound pressure into force, increasing that force through a lever system, and applying the increased force to a small area of the cochlea, thereby creating a pressure about 40 times that in the original sound wave. A protective muscle reaction to intense sounds greatly reduces the mechanical advantage of the lever system.

[link] shows the middle and inner ear in greater detail. Pressure waves moving through the cochlea cause the tectorial membrane to vibrate, rubbing cilia (called hair cells), which stimulate nerves that send electrical signals to the brain. The membrane resonates at different positions for different frequencies, with high frequencies stimulating nerves at the near end and low frequencies at the far end. The complete operation of the cochlea is still not understood, but several mechanisms for sending information to the brain are known to be involved. For sounds below about 1000 Hz, the nerves send signals at the same frequency as the sound. For frequencies greater than about 1000 Hz, the nerves signal frequency by position. There is a structure to the cilia, and there are connections between nerve cells that perform signal processing before information is sent to the brain. Intensity information is partly indicated by the number of nerve signals and by volleys of signals. The brain processes the cochlear nerve signals to provide additional information such as source direction (based on time and intensity comparisons of sounds from both ears). Higher-level processing produces many nuances, such as music appreciation.



The inner ear, or cochlea, is a coiled tube about 3 mm in diameter and 3 cm in length if uncoiled. When the oval window is forced inward, as shown, a pressure wave travels through the perilymph in the direction of the arrows, stimulating nerves at the base of cilia in the organ of Corti.

Hearing losses can occur because of problems in the middle or inner ear. Conductive losses in the middle ear can be partially overcome by sending sound vibrations to the cochlea through the skull. Hearing aids for this purpose usually press against the bone behind the ear, rather than simply amplifying the sound sent into the ear canal as many hearing aids do. Damage to the nerves in the cochlea is not repairable, but amplification can partially compensate. There is a risk that amplification will produce further damage. Another common failure in the cochlea is damage or loss of the cilia but with nerves remaining functional. Cochlear implants that stimulate the nerves directly are now available and widely accepted. Over 100,000 implants are in use, in about equal numbers of adults and children.

The cochlear implant was pioneered in Melbourne, Australia, by Graeme Clark in the 1970s for his deaf father. The implant consists of three external components and two internal components. The external components are a microphone for picking up sound and converting it into an electrical signal, a speech processor to select certain frequencies and a transmitter to transfer the signal to the internal components through electromagnetic induction. The internal components consist of a receiver/transmitter secured in the bone beneath the skin, which converts the signals into electric impulses and sends them through an internal cable to the cochlea and an array of about 24 electrodes wound through the cochlea. These electrodes in turn send the impulses directly into the brain. The electrodes basically emulate the cilia.

Exercise:

Check Your Understanding

Problem:

Are ultrasound and infrasound imperceptible to all hearing organisms? Explain your answer.

Solution:

No, the range of perceptible sound is based in the range of human hearing. Many other organisms perceive either infrasound or ultrasound.

Section Summary

- The range of audible frequencies is 20 to 20,000 Hz.
- Those sounds above 20,000 Hz are ultrasound, whereas those below 20 Hz are infrasound.
- The perception of frequency is pitch.
- The perception of intensity is loudness.
- Loudness has units of phons.

Conceptual Questions

Exercise:

Problem:

Why can a hearing test show that your threshold of hearing is 0 dB at 250 Hz, when [link] implies that no one can hear such a frequency at less than 20 dB?

Problems & Exercises

Exercise:

Problem:

The factor of 10^{-12} in the range of intensities to which the ear can respond, from threshold to that causing damage after brief exposure, is truly remarkable. If you could measure distances over the same range with a single instrument and the smallest distance you could measure was 1 mm, what would the largest be?

Solution:
Equation:

$$1 \times 10^{6}\text{ km}$$

Exercise:

Problem:

The frequencies to which the ear responds vary by a factor of 10^3. Suppose the speedometer on your car measured speeds differing by the same factor of 10^3, and the greatest speed it reads is 90.0 mi/h. What would be the slowest nonzero speed it could read?

Exercise:

Problem:

What are the closest frequencies to 500 Hz that an average person can clearly distinguish as being different in frequency from 500 Hz? The sounds are not present simultaneously.

Solution:

498.5 or 501.5 Hz

Exercise:

Problem:

Can the average person tell that a 2002-Hz sound has a different frequency than a 1999-Hz sound without playing them simultaneously?

Exercise:

Problem:

If your radio is producing an average sound intensity level of 85 dB, what is the next lowest sound intensity level that is clearly less intense?

Solution:

82 dB

Exercise:

Problem:

Can you tell that your roommate turned up the sound on the TV if its average sound intensity level goes from 70 to 73 dB?

Exercise:

Problem:

Based on the graph in [link], what is the threshold of hearing in decibels for frequencies of 60, 400, 1000, 4000, and 15,000 Hz? Note that many AC electrical appliances produce 60 Hz, music is commonly 400 Hz, a reference frequency is 1000 Hz, your maximum sensitivity is near 4000 Hz, and many older TVs produce a 15,750 Hz whine.

Solution:

approximately 48, 9, 0, –7, and 20 dB, respectively

Exercise:

Problem:

What sound intensity levels must sounds of frequencies 60, 3000, and 8000 Hz have in order to have the same loudness as a 40-dB sound of frequency 1000 Hz (that is, to have a loudness of 40 phons)?

Exercise:

Problem:

What is the approximate sound intensity level in decibels of a 600-Hz tone if it has a loudness of 20 phons? If it has a loudness of 70 phons?

Solution:

(a) 23 dB

(b) 70 dB

Exercise:

Problem:

(a) What are the loudnesses in phons of sounds having frequencies of 200, 1000, 5000, and 10,000 Hz, if they are all at the same 60.0-dB sound intensity level? (b) If they are all at 110 dB? (c) If they are all at 20.0 dB?

Exercise:

Problem:

Suppose a person has a 50-dB hearing loss at all frequencies. By how many factors of 10 will low-intensity sounds need to be amplified to seem normal to this person? Note that smaller amplification is appropriate for more intense sounds to avoid further hearing damage.

Solution:

Five factors of 10

Exercise:

Problem:

If a woman needs an amplification of 5.0×10^{12} times the threshold intensity to enable her to hear at all frequencies, what is her overall hearing loss in dB? Note that smaller amplification is appropriate for more intense sounds to avoid further damage to her hearing from levels above 90 dB.

Exercise:

Problem:

(a) What is the intensity in watts per meter squared of a just barely audible 200-Hz sound? (b) What is the intensity in watts per meter squared of a barely audible 4000-Hz sound?

Solution:

(a) $2 \times 10^{-10}\ \mathrm{W/m^2}$

(b) $2 \times 10^{-13}\ \mathrm{W/m^2}$

Exercise:

Problem:

(a) Find the intensity in watts per meter squared of a 60.0-Hz sound having a loudness of 60 phons. (b) Find the intensity in watts per meter squared of a 10,000-Hz sound having a loudness of 60 phons.

Exercise:

Problem:

A person has a hearing threshold 10 dB above normal at 100 Hz and 50 dB above normal at 4000 Hz. How much more intense must a 100-Hz tone be than a 4000-Hz tone if they are both barely audible to this person?

Solution:

2.5

Exercise:

Problem:

A child has a hearing loss of 60 dB near 5000 Hz, due to noise exposure, and normal hearing elsewhere. How much more intense is a 5000-Hz tone than a 400-Hz tone if they are both barely audible to the child?

Exercise:

Problem:

What is the ratio of intensities of two sounds of identical frequency if the first is just barely discernible as louder to a person than the second?

Solution:

1.26

Glossary

loudness
the perception of sound intensity

timbre
number and relative intensity of multiple sound frequencies

note
basic unit of music with specific names, combined to generate tunes

tone
number and relative intensity of multiple sound frequencies

phon
the numerical unit of loudness

ultrasound
sounds above 20,000 Hz

infrasound
sounds below 20 Hz

Ultrasound

- Define acoustic impedance and intensity reflection coefficient.
- Describe medical and other uses of ultrasound technology.
- Calculate acoustic impedance using density values and the speed of ultrasound.
- Calculate the velocity of a moving object using Doppler-shifted ultrasound.

Ultrasound is used in medicine to painlessly and noninvasively monitor patient health and diagnose a wide range of disorders. (credit: abbybatchelder, Flickr)

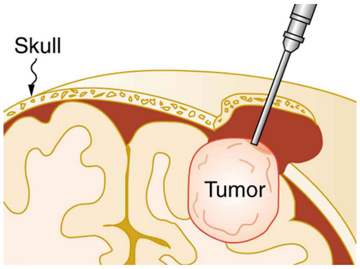
Any sound with a frequency above 20,000 Hz (or 20 kHz)—that is, above the highest audible frequency—is defined to be ultrasound. In practice, it is possible to create ultrasound frequencies up to more than a gigahertz. (Higher frequencies are difficult to create; furthermore, they propagate poorly because they are very strongly absorbed.) Ultrasound has a tremendous number of applications, which range from burglar alarms to use in cleaning delicate objects to the guidance systems of bats. We begin our discussion of ultrasound with some of its applications in medicine, in which it is used extensively both for diagnosis and for therapy.

Note:
Characteristics of Ultrasound
The characteristics of ultrasound, such as frequency and intensity, are wave properties common to all types of waves. Ultrasound also has a wavelength that limits the fineness of detail it can detect. This characteristic is true of all waves. We can never observe details significantly smaller than the wavelength of our probe; for example,

we will never see individual atoms with visible light, because the atoms are so small compared with the wavelength of light.

Ultrasound in Medical Therapy

Ultrasound, like any wave, carries energy that can be absorbed by the medium carrying it, producing effects that vary with intensity. When focused to intensities of 10^3 to $10^5\ \text{W/m}^2$, ultrasound can be used to shatter gallstones or pulverize cancerous tissue in surgical procedures. (See [link].) Intensities this great can damage individual cells, variously causing their protoplasm to stream inside them, altering their permeability, or rupturing their walls through *cavitation*. Cavitation is the creation of vapor cavities in a fluid—the longitudinal vibrations in ultrasound alternatively compress and expand the medium, and at sufficient amplitudes the expansion separates molecules. Most cavitation damage is done when the cavities collapse, producing even greater shock pressures.



The tip of this small probe oscillates at 23 kHz with such a large amplitude that it pulverizes tissue on contact. The debris is then aspirated. The speed of the tip may exceed the speed of sound in tissue, thus creating shock waves and cavitation, rather than a smooth

simple harmonic
oscillator–type
wave.

Most of the energy carried by high-intensity ultrasound in tissue is converted to thermal energy. In fact, intensities of 10^3 to $10^4\ \mathrm{W/m^2}$ are commonly used for deep-heat treatments called ultrasound diathermy. Frequencies of 0.8 to 1 MHz are typical. In both athletics and physical therapy, ultrasound diathermy is most often applied to injured or overworked muscles to relieve pain and improve flexibility. Skill is needed by the therapist to avoid “bone burns” and other tissue damage caused by overheating and cavitation, sometimes made worse by reflection and focusing of the ultrasound by joint and bone tissue.

In some instances, you may encounter a different decibel scale, called the sound *pressure* level, when ultrasound travels in water or in human and other biological tissues. We shall not use the scale here, but it is notable that numbers for sound pressure levels range 60 to 70 dB higher than you would quote for β, the sound intensity level used in this text. Should you encounter a sound pressure level of 220 decibels, then, it is not an astronomically high intensity, but equivalent to about 155 dB—high enough to destroy tissue, but not as unreasonably high as it might seem at first.

Ultrasound in Medical Diagnostics

When used for imaging, ultrasonic waves are emitted from a transducer, a crystal exhibiting the piezoelectric effect (the expansion and contraction of a substance when a voltage is applied across it, causing a vibration of the crystal). These high-frequency vibrations are transmitted into any tissue in contact with the transducer. Similarly, if a pressure is applied to the crystal (in the form of a wave reflected off tissue layers), a voltage is produced which can be recorded. The crystal therefore acts as both a transmitter and a receiver of sound. Ultrasound is also partially absorbed by tissue on its path, both on its journey away from the transducer and on its return journey. From the time between when the original signal is sent and when the reflections from various boundaries between media are received, (as well as a measure of the intensity loss of the signal), the nature and position of each boundary between tissues and organs may be deduced.

Reflections at boundaries between two different media occur because of differences in a characteristic known as the **acoustic impedance** Z of each substance. Impedance is defined as

Equation:

$$Z = \rho v,$$

where ρ is the density of the medium (in kg/m^3) and v is the speed of sound through the medium (in m/s). The units for Z are therefore $\text{kg/(m}^2 \cdot \text{s)}$.

[link] shows the density and speed of sound through various media (including various soft tissues) and the associated acoustic impedances. Note that the acoustic impedances for soft tissue do not vary much but that there is a big difference between the acoustic impedance of soft tissue and air and also between soft tissue and bone.

Medium	**Density (kg/m^3)**	**Speed of Ultrasound (m/s)**	**Acoustic Impedance** $\left(\text{kg}/\left(\text{m}^2 \cdot \text{s}\right)\right)$
Air	1.3	330	429
Water	1000	1500	1.5×10^6
Blood	1060	1570	1.66×10^6
Fat	925	1450	1.34×10^6
Muscle (average)	1075	1590	1.70×10^6
Bone (varies)	1400–1900	4080	5.7×10^6 to 7.8×10^6
Barium titanate (transducer material)	5600	5500	30.8×10^6

The Ultrasound Properties of Various Media, Including Soft Tissue Found in the Body

At the boundary between media of different acoustic impedances, some of the wave energy is reflected and some is transmitted. The greater the *difference* in acoustic impedance between the two media, the greater the reflection and the smaller the transmission.

The **intensity reflection coefficient** a is defined as the ratio of the intensity of the reflected wave relative to the incident (transmitted) wave. This statement can be written mathematically as

Equation:

$$a = \frac{\left(Z_2 - Z_1\right)^2}{\left(Z_1 + Z_2\right)^2},$$

where Z_1 and Z_2 are the acoustic impedances of the two media making up the boundary. A reflection coefficient of zero (corresponding to total transmission and no reflection) occurs when the acoustic impedances of the two media are the same. An impedance "match" (no reflection) provides an efficient coupling of sound energy from one medium to another. The image formed in an ultrasound is made by tracking reflections (as shown in [link]) and mapping the intensity of the reflected sound waves in a two-dimensional plane.

Example:

Calculate Acoustic Impedance and Intensity Reflection Coefficient: Ultrasound and Fat Tissue

(a) Using the values for density and the speed of ultrasound given in [link], show that the acoustic impedance of fat tissue is indeed $1.34 \times 10^6 \text{ kg/(m}^2\text{·s)}$.

(b) Calculate the intensity reflection coefficient of ultrasound when going from fat to muscle tissue.

Strategy for (a)

The acoustic impedance can be calculated using $Z = \text{ρv}$ and the values for ρ and v found in [link].

Solution for (a)

(1) Substitute known values from [link] into $Z = \text{ρv}$.

Equation:

$$Z = \rho v = \left(925 \text{ kg/m}^3\right)(1450 \text{ m/s})$$

(2) Calculate to find the acoustic impedance of fat tissue.

Equation:

$$1.34 \times 10^6 \text{ kg/(m}^2\text{·s)}$$

This value is the same as the value given for the acoustic impedance of fat tissue.

Strategy for (b)

The intensity reflection coefficient for any boundary between two media is given by $a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$, and the acoustic impedance of muscle is given in [link].

Solution for (b)

Substitute known values into $a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2}$ to find the intensity reflection coefficient:

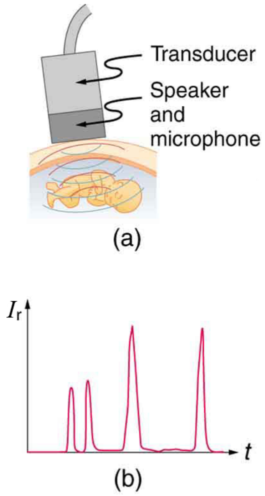
Equation:

$$a = \frac{(Z_2 - Z_1)^2}{(Z_1 + Z_2)^2} = \frac{\left(1.34 \times 10^6 \text{ kg/(m}^2\text{· s)} - 1.70 \times 10^6 \text{ kg/(m}^2\text{· s)}\right)^2}{\left(1.70 \times 10^6 \text{ kg/(m}^2\text{· s)} + 1.34 \times 10^6 \text{ kg/(m}^2\text{· s)}\right)^2} = 0.014$$

Discussion

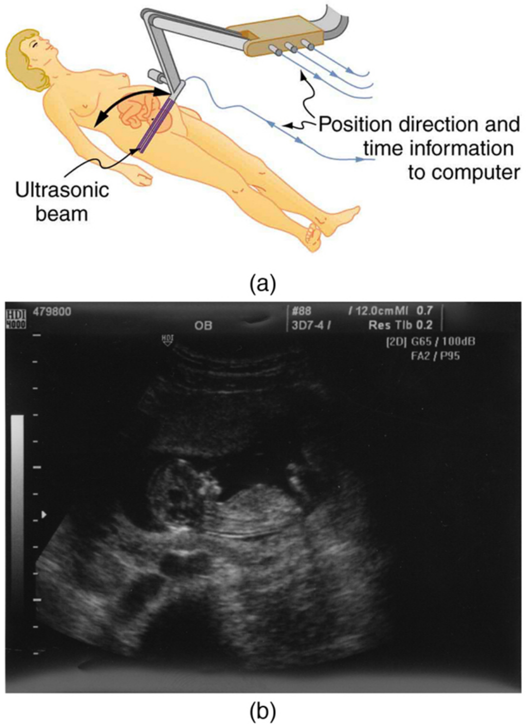
This result means that only 1.4% of the incident intensity is reflected, with the remaining being transmitted.

The applications of ultrasound in medical diagnostics have produced untold benefits with no known risks. Diagnostic intensities are too low (about 10^{-2} W/m^2) to cause thermal damage. More significantly, ultrasound has been in use for several decades and detailed follow-up studies do not show evidence of ill effects, quite unlike the case for x-rays.



(a) An ultrasound speaker doubles as a microphone. Brief bleeps are broadcast, and echoes are recorded from various depths. (b) Graph of echo intensity versus time. The time for echoes to return is directly proportional to the distance of the reflector, yielding this information noninvasively.

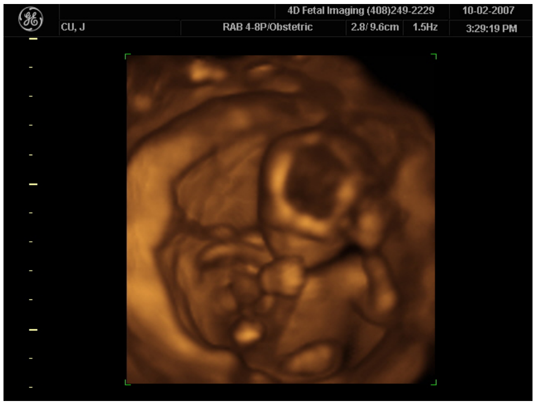
The most common ultrasound applications produce an image like that shown in [link]. The speaker-microphone broadcasts a directional beam, sweeping the beam across the area of interest. This is accomplished by having multiple ultrasound sources in the probe’s head, which are phased to interfere constructively in a given, adjustable direction. Echoes are measured as a function of position as well as depth. A computer constructs an image that reveals the shape and density of internal structures.



(a) An ultrasonic image is produced by sweeping the ultrasonic beam across the area of interest, in this case the woman’s abdomen. Data are recorded and analyzed in a computer, providing a two-dimensional image. (b) Ultrasound image of 12-week-old fetus. (credit: Margaret W. Carruthers, Flickr)

How much detail can ultrasound reveal? The image in [link] is typical of low-cost systems, but that in [link] shows the remarkable detail possible with more advanced systems, including 3D imaging. Ultrasound today is commonly used in prenatal care. Such imaging can be used to see if the fetus is developing at a normal rate, and help in the determination of serious problems early in the pregnancy. Ultrasound is also in wide use to image the chambers of the heart and the flow of blood within the beating heart, using the Doppler effect (echocardiology).

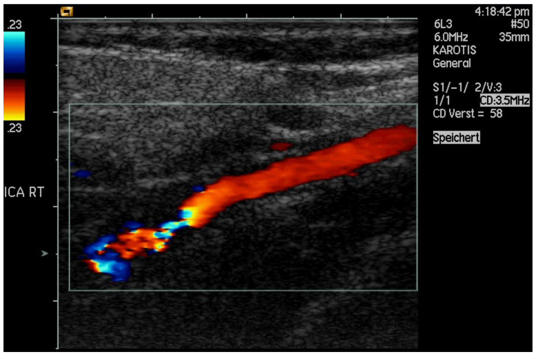
Whenever a wave is used as a probe, it is very difficult to detect details smaller than its wavelength λ. Indeed, current technology cannot do quite this well. Abdominal scans may use a 7-MHz frequency, and the speed of sound in tissue is about 1540 m/s —so the wavelength limit to detail would be $\lambda = \frac{v_\text{w}}{f} = \frac{1540\text{ m/s}}{7\times10^6\text{ Hz}} = 0.22\text{ mm}$. In practice, 1-mm detail is attainable, which is sufficient for many purposes. Higher-frequency ultrasound would allow greater detail, but it does not penetrate as well as lower frequencies do. The accepted rule of thumb is that you can effectively scan to a depth of about 500λ into tissue. For 7 MHz, this penetration limit is $500 \times 0.22\text{ mm}$, which is 0.11 m. Higher frequencies may be employed in smaller organs, such as the eye, but are not practical for looking deep into the body.



A 3D ultrasound image of a fetus. As well as for the detection of any abnormalities, such scans have also been shown to be useful for strengthening the emotional bonding between parents and their unborn child. (credit: Jennie Cu, Wikimedia Commons)

In addition to shape information, ultrasonic scans can produce density information superior to that found in X-rays, because the intensity of a reflected sound is related to changes in density. Sound is most strongly reflected at places where density changes are greatest.

Another major use of ultrasound in medical diagnostics is to detect motion and determine velocity through the Doppler shift of an echo, known as **Doppler-shifted ultrasound**. This technique is used to monitor fetal heartbeat, measure blood velocity, and detect occlusions in blood vessels, for example. (See [link].) The magnitude of the Doppler shift in an echo is directly proportional to the velocity of whatever reflects the sound. Because an echo is involved, there is actually a double shift. The first occurs because the reflector (say a fetal heart) is a moving observer and receives a Doppler-shifted frequency. The reflector then acts as a moving source, producing a second Doppler shift.



This Doppler-shifted ultrasonic image of a partially occluded artery uses color to indicate velocity. The highest velocities are in red, while the lowest are blue. The blood must move faster through the constriction to carry the same flow. (credit: Arning C, Grzyska U, Wikimedia Commons)

A clever technique is used to measure the Doppler shift in an echo. The frequency of the echoed sound is superimposed on the broadcast frequency, producing beats. The beat frequency is $F_{\text{B}} = \mid f_1 - f_2 \mid$, and so it is directly proportional to the Doppler shift $(f_1 - f_2)$ and hence, the reflector's velocity. The advantage in this technique is that the Doppler shift is small (because the reflector's velocity is small), so that great accuracy would be needed to measure the shift directly. But measuring the beat frequency is easy, and it is not affected if the broadcast frequency varies somewhat. Furthermore, the beat frequency is in the audible range and can be amplified for audio feedback to the medical observer.

Note:

Uses for Doppler-Shifted Radar

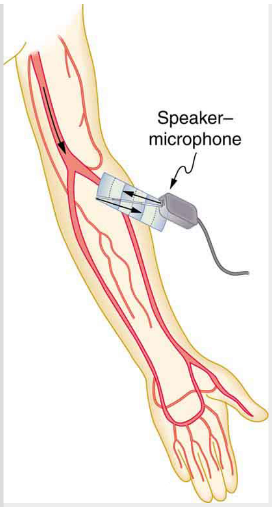
Doppler-shifted radar echoes are used to measure wind velocities in storms as well as aircraft and automobile speeds. The principle is the same as for Doppler-shifted ultrasound. There is evidence that bats and dolphins may also sense the velocity of an object (such as prey) reflecting their ultrasound signals by observing its Doppler shift.

Example:

Calculate Velocity of Blood: Doppler-Shifted Ultrasound

Ultrasound that has a frequency of 2.50 MHz is sent toward blood in an artery that is moving toward the source at 20.0 cm/s, as illustrated in [link]. Use the speed of sound in human tissue as 1540 m/s. (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

a. What frequency does the blood receive?
b. What frequency returns to the source?
c. What beat frequency is produced if the source and returning frequencies are mixed?



Ultrasound is partly reflected by blood cells and plasma back toward the speaker-microphone. Because the cells are moving, two Doppler shifts are produced—one for blood as a moving observer, and the other for the reflected sound coming from a moving source. The magnitude of the shift is directly proportional to blood velocity.

Strategy

The first two questions can be answered using $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} \pm v_{\text{s}}}\right)$ and $f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right)$ for the Doppler shift. The last question asks for beat frequency, which is the difference between the original and returning frequencies.

Solution for (a)

(1) Identify knowns:

- The blood is a moving observer, and so the frequency it receives is given by
  Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}} \pm v_{\text{obs}}}{v_{\text{w}}}\right).$$

- v_{b} is the blood velocity (v_{obs} here) and the plus sign is chosen because the motion is toward the source.

(2) Enter the given values into the equation.

Equation:

$$f_{\text{obs}} = (2{,}500{,}000\text{ Hz})\left(\frac{1540\text{ m/s} + 0.2\text{ m/s}}{1540\text{ m/s}}\right)$$

(3) Calculate to find the frequency: 2,500,325 Hz.

Solution for (b)

(1) Identify knowns:

- The blood acts as a moving source.
- The microphone acts as a stationary observer.
- The frequency leaving the blood is 2,500,325 Hz, but it is shifted upward as given by
  Equation:

$$f_{\text{obs}} = f_{\text{s}}\left(\frac{v_{\text{w}}}{v_{\text{w}} - v_{\text{b}}}\right).$$

  f_{obs} is the frequency received by the speaker-microphone.
- The source velocity is v_{b}.
- The minus sign is used because the motion is toward the observer.

The minus sign is used because the motion is toward the observer.
(2) Enter the given values into the equation:
Equation:

$$f_{\text{obs}} = (2{,}500{,}325\text{ Hz})\left(\frac{1540\text{ m/s}}{1540\text{ m/s} - 0.200\text{ m/s}}\right)$$

(3) Calculate to find the frequency returning to the source: 2,500,649 Hz.
Solution for (c)
(1) Identify knowns:

- The beat frequency is simply the absolute value of the difference between f_{s} and f_{obs}, as stated in:
  Equation:

  $$f_{\text{B}} = \mid f_{\text{obs}} - f_{\text{s}} \mid.$$

(2) Substitute known values:
Equation:

$$\mid 2{,}500{,}649\text{ Hz} - 2{,}500{,}000\text{ Hz} \mid$$

(3) Calculate to find the beat frequency: 649 Hz.
Discussion
The Doppler shifts are quite small compared with the original frequency of 2.50 MHz. It is far easier to measure the beat frequency than it is to measure the echo frequency with an accuracy great enough to see shifts of a few hundred hertz out of a couple of megahertz. Furthermore, variations in the source frequency do not greatly affect the beat frequency, because both f_{s} and f_{obs}would increase or decrease. Those changes subtract out in $f_{\text{B}} = \mid f_{\text{obs}} - f_{\text{s}} \mid$.

Note:
Industrial and Other Applications of Ultrasound
Industrial, retail, and research applications of ultrasound are common. A few are discussed here. Ultrasonic cleaners have many uses. Jewelry, machined parts, and other objects that have odd shapes and crevices are immersed in a cleaning fluid that is agitated with ultrasound typically about 40 kHz in frequency. The intensity is great enough to cause cavitation, which is responsible for most of the cleansing action. Because cavitation-produced shock pressures are large and well transmitted in a fluid,

they reach into small crevices where even a low-surface-tension cleaning fluid might not penetrate.

Sonar is a familiar application of ultrasound. Sonar typically employs ultrasonic frequencies in the range from 30.0 to 100 kHz. Bats, dolphins, submarines, and even some birds use ultrasonic sonar. Echoes are analyzed to give distance and size information both for guidance and finding prey. In most sonar applications, the sound reflects quite well because the objects of interest have significantly different density than the medium in which they travel. When the Doppler shift is observed, velocity information can also be obtained. Submarine sonar can be used to obtain such information, and there is evidence that some bats also sense velocity from their echoes.

Similarly, there are a range of relatively inexpensive devices that measure distance by timing ultrasonic echoes. Many cameras, for example, use such information to focus automatically. Some doors open when their ultrasonic ranging devices detect a nearby object, and certain home security lights turn on when their ultrasonic rangers observe motion. Ultrasonic “measuring tapes” also exist to measure such things as room dimensions. Sinks in public restrooms are sometimes automated with ultrasound devices to turn faucets on and off when people wash their hands. These devices reduce the spread of germs and can conserve water.

Ultrasound is used for nondestructive testing in industry and by the military. Because ultrasound reflects well from any large change in density, it can reveal cracks and voids in solids, such as aircraft wings, that are too small to be seen with x-rays. For similar reasons, ultrasound is also good for measuring the thickness of coatings, particularly where there are several layers involved.

Basic research in solid state physics employs ultrasound. Its attenuation is related to a number of physical characteristics, making it a useful probe. Among these characteristics are structural changes such as those found in liquid crystals, the transition of a material to a superconducting phase, as well as density and other properties.

These examples of the uses of ultrasound are meant to whet the appetites of the curious, as well as to illustrate the underlying physics of ultrasound. There are many more applications, as you can easily discover for yourself.

Exercise:
Check Your Understanding

Problem:

Why is it possible to use ultrasound both to observe a fetus in the womb and also to destroy cancerous tumors in the body?

Solution:

Ultrasound can be used medically at different intensities. Lower intensities do not cause damage and are used for medical imaging. Higher intensities can pulverize and destroy targeted substances in the body, such as tumors.

Section Summary

- The acoustic impedance is defined as:
  Equation:

  $$Z=\rho v,$$

  ρ is the density of a medium through which the sound travels and v is the speed of sound through that medium.
- The intensity reflection coefficient a, a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave, is given by
  Equation:

  $$a=\frac{{(Z_2-Z_1)}^2}{{(Z_1+Z_2)}^2}.$$

- The intensity reflection coefficient is a unitless quantity.

Conceptual Questions

Exercise:

Problem:

If audible sound follows a rule of thumb similar to that for ultrasound, in terms of its absorption, would you expect the high or low frequencies from your neighbor’s stereo to penetrate into your house? How does this expectation compare with your experience?

Exercise:

Problem:

Elephants and whales are known to use infrasound to communicate over very large distances. What are the advantages of infrasound for long distance communication?

Exercise:

Problem:

It is more difficult to obtain a high-resolution ultrasound image in the abdominal region of someone who is overweight than for someone who has a slight build. Explain why this statement is accurate.

Exercise:

Problem:

Suppose you read that 210-dB ultrasound is being used to pulverize cancerous tumors. You calculate the intensity in watts per centimeter squared and find it is unreasonably high ($10^5\ \mathrm{W/cm^2}$). What is a possible explanation?

Problems & Exercises

Unless otherwise indicated, for problems in this section, assume that the speed of sound through human tissues is 1540 m/s.

Exercise:

Problem:

What is the sound intensity level in decibels of ultrasound of intensity $10^5\ \mathrm{W/m^2}$, used to pulverize tissue during surgery?

Solution:

170 dB

Exercise:

Problem:

Is 155-dB ultrasound in the range of intensities used for deep heating? Calculate the intensity of this ultrasound and compare this intensity with values quoted in the text.

Exercise:

Problem:

Find the sound intensity level in decibels of $2.00 \times 10^{-2}\ \mathrm{W/m^2}$ ultrasound used in medical diagnostics.

Solution:

103 dB

Exercise:

Problem:

The time delay between transmission and the arrival of the reflected wave of a signal using ultrasound traveling through a piece of fat tissue was 0.13 ms. At what depth did this reflection occur?

Exercise:

Problem:

In the clinical use of ultrasound, transducers are always coupled to the skin by a thin layer of gel or oil, replacing the air that would otherwise exist between the transducer and the skin. (a) Using the values of acoustic impedance given in [link] calculate the intensity reflection coefficient between transducer material and air. (b) Calculate the intensity reflection coefficient between transducer material and gel (assuming for this problem that its acoustic impedance is identical to that of water). (c) Based on the results of your calculations, explain why the gel is used.

Solution:

(a) 1.00

(b) 0.823

(c) Gel is used to facilitate the transmission of the ultrasound between the transducer and the patient’s body.

Exercise:

Problem:

(a) Calculate the minimum frequency of ultrasound that will allow you to see details as small as 0.250 mm in human tissue. (b) What is the effective depth to which this sound is effective as a diagnostic probe?

Exercise:

Problem:

(a) Find the size of the smallest detail observable in human tissue with 20.0-MHz ultrasound. (b) Is its effective penetration depth great enough to examine the entire eye (about 3.00 cm is needed)? (c) What is the wavelength of such ultrasound in 0°C air?

Solution:

(a) $77.0\ \mu\text{m}$

(b) Effective penetration depth = 3.85 cm, which is enough to examine the eye.

(c) $16.6\ \mu\text{m}$

Exercise:

Problem:

(a) Echo times are measured by diagnostic ultrasound scanners to determine distances to reflecting surfaces in a patient. What is the difference in echo times for tissues that are 3.50 and 3.60 cm beneath the surface? (This difference is the minimum resolving time for the scanner to see details as small as 0.100 cm, or 1.00 mm. Discrimination of smaller time differences is needed to see smaller details.) (b) Discuss whether the period T of this ultrasound must be smaller than the minimum time resolution. If so, what is the minimum frequency of the ultrasound and is that out of the normal range for diagnostic ultrasound?

Exercise:

Problem:

(a) How far apart are two layers of tissue that produce echoes having round-trip times (used to measure distances) that differ by $0.750\ \mu\text{s}$? (b) What minimum frequency must the ultrasound have to see detail this small?

Solution:

(a) $5.78 \times 10^{-4}\ \text{m}$

(b) $2.67 \times 10^{6}\ \text{Hz}$

Exercise:

Problem:

(a) A bat uses ultrasound to find its way among trees. If this bat can detect echoes 1.00 ms apart, what minimum distance between objects can it detect? (b) Could this distance explain the difficulty that bats have finding an open door when they accidentally get into a house?

Exercise:

Problem:

A dolphin is able to tell in the dark that the ultrasound echoes received from two sharks come from two different objects only if the sharks are separated by 3.50 m, one being that much farther away than the other. (a) If the ultrasound has a frequency of 100 kHz, show this ability is not limited by its wavelength. (b) If this ability is due to the dolphin’s ability to detect the arrival times of echoes, what is the minimum time difference the dolphin can perceive?

Solution:

(a) $v_{\text{w}} = 1540\ \text{m/s} = f\lambda \Rightarrow \lambda = \frac{1540\ \text{m/s}}{100 \times 10^{3}\ \text{Hz}} = 0.0154\ \text{m} < 3.50\ \text{m}$. Because the wavelength is much shorter than the distance in question, the wavelength is not the limiting factor.

(b) 4.55 ms

Exercise:

Problem:

A diagnostic ultrasound echo is reflected from moving blood and returns with a frequency 500 Hz higher than its original 2.00 MHz. What is the velocity of the blood? (Assume that the frequency of 2.00 MHz is accurate to seven significant figures and 500 Hz is accurate to three significant figures.)

Exercise:

Problem:

Ultrasound reflected from an oncoming bloodstream that is moving at 30.0 cm/s is mixed with the original frequency of 2.50 MHz to produce beats. What is the beat frequency? (Assume that the frequency of 2.50 MHz is accurate to seven significant figures.)

Solution:

974 Hz

(Note: extra digits were retained in order to show the difference.)

Glossary

acoustic impedance
property of medium that makes the propagation of sound waves more difficult

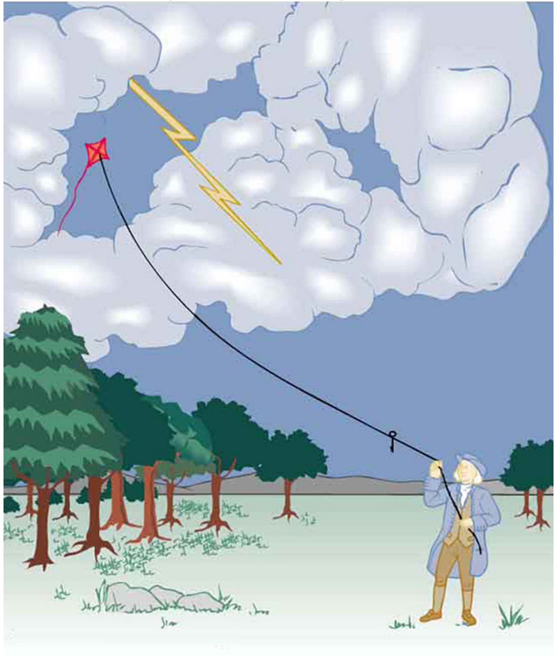
intensity reflection coefficient
a measure of the ratio of the intensity of the wave reflected off a boundary between two media relative to the intensity of the incident wave

Doppler-shifted ultrasound
a medical technique to detect motion and determine velocity through the Doppler shift of an echo

Introduction to Electric Charge and Electric Field class="introduction"

Static electricity from this plastic slide causes the child’s hair to stand on end. The sliding motion stripped electrons away from the child’s body, leaving an excess of positive charges, which repel each other along each strand of hair. (credit: Ken Bosma/Wikimedia Commons)

The image of American politician and scientist Benjamin Franklin (1706–1790) flying a kite in a thunderstorm is familiar to every schoolchild. (See [link].) In this experiment, Franklin demonstrated a connection between lightning and **static electricity**. Sparks were drawn from a key hung on a kite string during an electrical storm. These sparks were like those produced by static electricity, such as the spark that jumps from your finger to a metal doorknob after you walk across a wool carpet. What Franklin demonstrated in his dangerous experiment was a connection between phenomena on two different scales: one the grand power of an electrical storm, the other an effect of more human proportions. Connections like this one reveal the underlying unity of the laws of nature, an aspect we humans find particularly appealing.



When Benjamin Franklin demonstrated that lightning was related to static electricity, he made a connection that is now part of the evidence that all directly experienced forces except the gravitational force are manifestations of the electromagnetic force.

Much has been written about Franklin. His experiments were only part of the life of a man who was a scientist, inventor, revolutionary, statesman, and writer. Franklin’s experiments were not performed in isolation, nor were they the only ones to reveal connections.

For example, the Italian scientist Luigi Galvani (1737–1798) performed a series of experiments in which static electricity was used to stimulate contractions of leg muscles of dead frogs, an effect already known in humans subjected to static discharges. But Galvani also found that if he joined two metal wires (say copper and zinc) end to end and touched the other ends to muscles, he produced the same effect in frogs as static discharge. Alessandro Volta (1745–1827), partly inspired by Galvani’s work, experimented with various combinations of metals and developed the battery.

During the same era, other scientists made progress in discovering fundamental connections. The periodic table was developed as the systematic properties of the elements were discovered. This influenced the development and refinement of the concept of atoms as the basis of matter. Such submicroscopic descriptions of matter also help explain a great deal more.

Atomic and molecular interactions, such as the forces of friction, cohesion, and adhesion, are now known to be manifestations of the **electromagnetic force**. Static electricity is just one aspect of the electromagnetic force, which also includes moving electricity and magnetism.

All the macroscopic forces that we experience directly, such as the sensations of touch and the tension in a rope, are due to the electromagnetic force, one of the four fundamental forces in nature. The gravitational force, another fundamental force, is actually sensed through the electromagnetic interaction of molecules, such as between those in our feet and those on the top of a bathroom scale. (The other two fundamental forces, the strong nuclear force and the weak nuclear force, cannot be sensed on the human scale.)

This chapter begins the study of electromagnetic phenomena at a fundamental level. The next several chapters will cover static electricity, moving electricity, and magnetism—collectively known as electromagnetism. In this chapter, we begin with the study of electric phenomena due to charges that are at least temporarily stationary, called electrostatics, or static electricity.

Glossary

static electricity
: a buildup of electric charge on the surface of an object

electromagnetic force
: one of the four fundamental forces of nature; the electromagnetic force consists of static electricity, moving electricity and magnetism

Static Electricity and Charge: Conservation of Charge

- Define electric charge, and describe how the two types of charge interact.
- Describe three common situations that generate static electricity.
- State the law of conservation of charge.

Borneo amber was mined in Sabah, Malaysia, from shale-sandstone-mudstone veins. When a piece of amber is rubbed with a piece of silk, the amber gains more electrons, giving it a net negative charge. At the same time, the silk, having lost electrons, becomes positively charged. (credit: Sebakoamber, Wikimedia Commons)

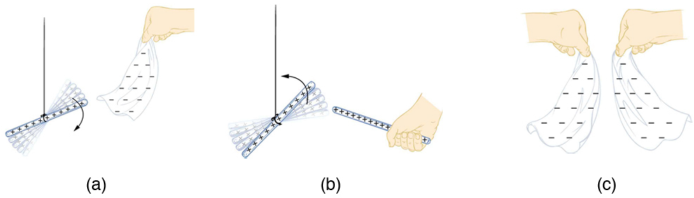
What makes plastic wrap cling? Static electricity. Not only are applications of static electricity common these days, its existence has been known since ancient times. The first record of its effects dates to ancient Greeks who noted more than 500 years B.C. that polishing amber temporarily enabled it to attract bits of straw (see [link]). The very word *electric* derives from the Greek word for amber (*electron*).

Many of the characteristics of static electricity can be explored by rubbing things together. Rubbing creates the spark you get from walking across a wool carpet, for example. Static cling generated in a clothes dryer and the attraction of straw to recently polished amber also result from rubbing. Similarly, lightning results from air movements under certain weather conditions. You can also rub a balloon on your hair, and the static electricity created can then make the balloon cling to a wall. We also have to be cautious of static electricity, especially in dry climates. When we pump gasoline, we are warned to discharge ourselves (after sliding across the seat) on a metal surface before grabbing the gas nozzle. Attendants in hospital operating rooms must wear booties with aluminum foil on the bottoms to avoid creating sparks which may ignite the oxygen being used.

Some of the most basic characteristics of static electricity include:

- The effects of static electricity are explained by a physical quantity not previously introduced, called electric charge.
- There are only two types of charge, one called positive and the other called negative.
- Like charges repel, whereas unlike charges attract.
- The force between charges decreases with distance.

How do we know there are two types of **electric charge**? When various materials are rubbed together in controlled ways, certain combinations of materials always produce one type of charge on one material and the opposite type on the other. By convention, we call one type of charge “positive”, and the other type “negative.” For example, when glass is rubbed with silk, the glass becomes positively charged and the silk negatively charged. Since the glass and silk have opposite charges, they attract one another like clothes that have rubbed together in a dryer. Two glass rods rubbed with silk in this manner will repel one another, since each rod has positive charge on it. Similarly, two silk cloths so rubbed will repel, since both cloths have negative charge. [link] shows how these simple materials can be used to explore the nature of the force between charges.



A glass rod becomes positively charged when rubbed with silk, while the silk becomes negatively charged. (a) The glass rod is attracted to the silk because their charges are opposite. (b) Two similarly charged glass rods repel. (c) Two similarly charged silk cloths repel.

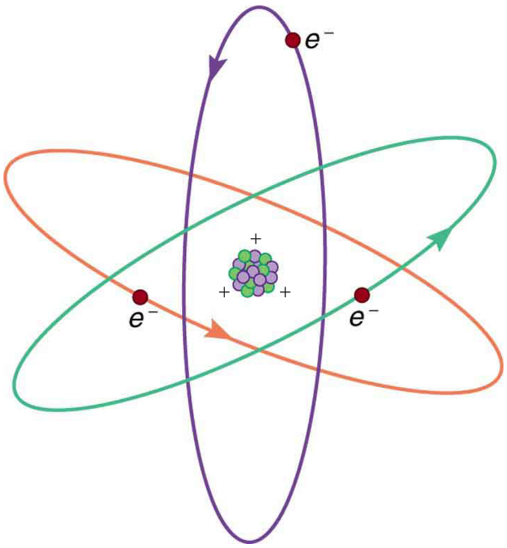
More sophisticated questions arise. Where do these charges come from? Can you create or destroy charge? Is there a smallest unit of charge? Exactly how does the force depend on the amount of charge and the distance between charges? Such questions obviously occurred to Benjamin Franklin and other early researchers, and they interest us even today.

Charge Carried by Electrons and Protons

Franklin wrote in his letters and books that he could see the effects of electric charge but did not understand what caused the phenomenon. Today we have the advantage of knowing that normal matter is made of atoms, and that atoms contain positive and negative charges, usually in equal amounts.

[link] shows a simple model of an atom with negative **electrons** orbiting its positive nucleus. The nucleus is positive due to the presence of positively charged **protons**. Nearly all charge in nature is due to electrons and protons, which are two of the three building blocks of most matter. (The third is the neutron, which is neutral, carrying no charge.) Other charge-carrying particles are observed in cosmic rays and nuclear decay, and are created in

particle accelerators. All but the electron and proton survive only a short time and are quite rare by comparison.



This simplified (and not to scale) view of an atom is called the planetary model of the atom. Negative electrons orbit a much heavier positive nucleus, as the planets orbit the much heavier sun. There the similarity ends, because forces in the atom are electromagnetic, whereas those in the planetary system are gravitational. Normal macroscopic amounts of matter contain immense numbers of atoms and molecules and, hence, even greater numbers of individual

negative and positive
charges.

The charges of electrons and protons are identical in magnitude but opposite in sign. Furthermore, all charged objects in nature are integral multiples of this basic quantity of charge, meaning that all charges are made of combinations of a basic unit of charge. Usually, charges are formed by combinations of electrons and protons. The magnitude of this basic charge is

Equation:

$$\mid q_e \mid = 1.60 \times 10^{-19}\ \text{C}.$$

The symbol q is commonly used for charge and the subscript e indicates the charge of a single electron (or proton).

The SI unit of charge is the coulomb (C). The number of protons needed to make a charge of 1.00 C is

Equation:

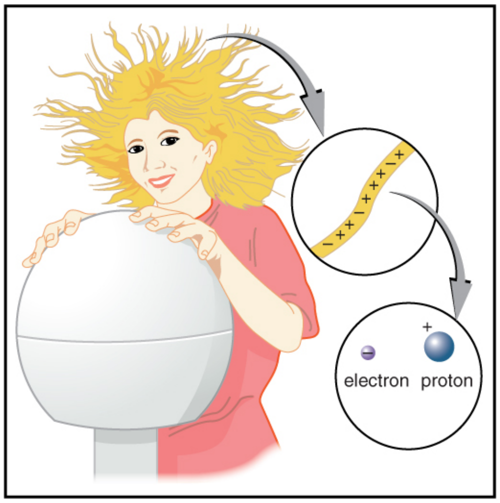
$$1.00\ \text{C} \times \frac{1\ \text{proton}}{1.60 \times 10^{-19}\ \text{C}} = 6.25 \times 10^{18}\ \text{protons}.$$

Similarly, 6.25×10^{18} electrons have a combined charge of −1.00 coulomb. Just as there is a smallest bit of an element (an atom), there is a smallest bit of charge. There is no directly observed charge smaller than $\mid q_e \mid$ (see Things Great and Small: The Submicroscopic Origin of Charge), and all observed charges are integral multiples of $\mid q_e \mid$.

Note:
Things Great and Small: The Submicroscopic Origin of Charge

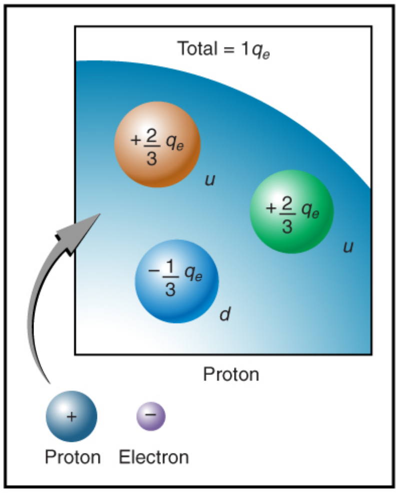
With the exception of exotic, short-lived particles, all charge in nature is carried by electrons and protons. Electrons carry the charge we have named negative. Protons carry an equal-magnitude charge that we call positive. (See [link].) Electron and proton charges are considered fundamental building blocks, since all other charges are integral multiples of those carried by electrons and protons. Electrons and protons are also two of the three fundamental building blocks of ordinary matter. The neutron is the third and has zero total charge.

[link] shows a person touching a Van de Graaff generator and receiving excess positive charge. The expanded view of a hair shows the existence of both types of charges but an excess of positive. The repulsion of these positive like charges causes the strands of hair to repel other strands of hair and to stand up. The further blowup shows an artist’s conception of an electron and a proton perhaps found in an atom in a strand of hair.



When this person touches a Van de Graaff generator, she receives an excess of positive charge, causing her hair to stand on end. The charges in one hair are shown. An artist’s conception of an electron and a proton illustrate the particles carrying the negative and positive charges. We cannot really see these particles with visible light because they are so small (the electron seems to be an infinitesimal point), but we know a great deal about their measurable properties, such as the charges they carry.

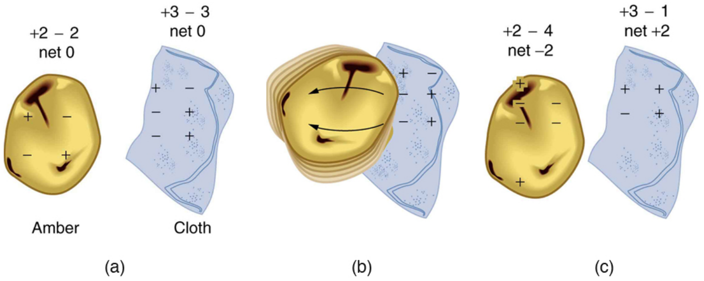
The electron seems to have no substructure; in contrast, when the substructure of protons is explored by scattering extremely energetic electrons from them, it appears that there are point-like particles inside the proton. These sub-particles, named quarks, have never been directly observed, but they are believed to carry fractional charges as seen in [link]. Charges on electrons and protons and all other directly observable particles are unitary, but these quark substructures carry charges of either $-\frac{1}{3}$ or $+\frac{2}{3}$. There are continuing attempts to observe fractional charge directly and to learn of the properties of quarks, which are perhaps the ultimate substructure of matter.



Artist’s conception of fractional quark charges inside a proton. A group of three quark charges add up to the single positive charge on the proton: $-\frac{1}{3}q_e + \frac{2}{3}q_e + \frac{2}{3}q_e = +1q_e$.

Separation of Charge in Atoms

Charges in atoms and molecules can be separated—for example, by rubbing materials together. Some atoms and molecules have a greater affinity for electrons than others and will become negatively charged by close contact in rubbing, leaving the other material positively charged. (See [link].) Positive charge can similarly be induced by rubbing. Methods other than rubbing can also separate charges. Batteries, for example, use combinations of substances that interact in such a way as to separate charges. Chemical interactions may transfer negative charge from one substance to the other, making one battery terminal negative and leaving the first one positive.



When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and only a few of them are shown here. (b) When rubbed together, some negative charge is transferred to the amber, leaving the cloth with a net positive charge. (c) When separated, the amber and cloth now have net charges, but the absolute value of the net positive and negative charges will be equal.

No charge is actually created or destroyed when charges are separated as we have been discussing. Rather, existing charges are moved about. In fact, in all situations the total amount of charge is always constant. This universally obeyed law of nature is called the **law of conservation of charge**.

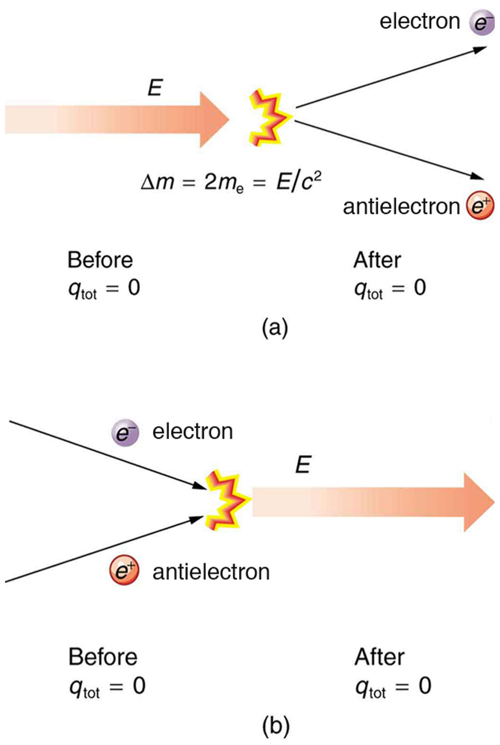
Note:
Law of Conservation of Charge
Total charge is constant in any process.

In more exotic situations, such as in particle accelerators, mass, Δm, can be created from energy in the amount $\Delta m = \frac{E}{c^2}$. Sometimes, the created mass is charged, such as when an electron is created. Whenever a charged particle is created, another having an opposite charge is always created along with it, so that the total charge created is zero. Usually, the two particles are “matter-antimatter” counterparts. For example, an antielectron would usually be created at the same time as an electron. The antielectron has a positive charge (it is called a positron), and so the total charge created is zero. (See [link].) All particles have antimatter counterparts with opposite signs. When matter and antimatter counterparts are brought together, they completely annihilate one another. By annihilate, we mean that the mass of the two particles is converted to energy E, again obeying the relationship $\Delta m = \frac{E}{c^2}$. Since the two particles have equal and opposite charge, the total charge is zero before and after the annihilation; thus, total charge is conserved.

Note:

Making Connections: Conservation Laws

Only a limited number of physical quantities are universally conserved. Charge is one—energy, momentum, and angular momentum are others. Because they are conserved, these physical quantities are used to explain more phenomena and form more connections than other, less basic quantities. We find that conserved quantities give us great insight into the rules followed by nature and hints to the organization of nature. Discoveries of conservation laws have led to further discoveries, such as the weak nuclear force and the quark substructure of protons and other particles.



(a) When enough energy is present, it can be converted into matter. Here the matter created is an electron–antielectron pair. (m_e is the electron's mass.) The total charge before and after this event is zero. (b) When matter and antimatter collide, they annihilate each other; the total charge is conserved at zero before and after the annihilation.

The law of conservation of charge is absolute—it has never been observed to be violated. Charge, then, is a special physical quantity, joining a very short list of other quantities in nature that are always conserved. Other conserved quantities include energy, momentum, and angular momentum.

Note:
PhET Explorations: Balloons and Static Electricity
Why does a balloon stick to your sweater? Rub a balloon on a sweater, then let go of the balloon and it flies over and sticks to the sweater. View the charges in the sweater, balloons, and the wall.
https://phet.colorado.edu/sims/html/balloons-and-static-electricity/latest/balloons-and-static-electricity_en.html

Section Summary

- There are only two types of charge, which we call positive and negative.
- Like charges repel, unlike charges attract, and the force between charges decreases with the square of the distance.
- The vast majority of positive charge in nature is carried by protons, while the vast majority of negative charge is carried by electrons.
- The electric charge of one electron is equal in magnitude and opposite in sign to the charge of one proton.
- An ion is an atom or molecule that has nonzero total charge due to having unequal numbers of electrons and protons.
- The SI unit for charge is the coulomb (C), with protons and electrons having charges of opposite sign but equal magnitude; the magnitude of this basic charge $\mid q_e \mid$ is
  Equation:

  $$\mid q_e \mid = 1.60 \times 10^{-19}\text{ C}.$$

- Whenever charge is created or destroyed, equal amounts of positive and negative are involved.
- Most often, existing charges are separated from neutral objects to obtain some net charge.

- Both positive and negative charges exist in neutral objects and can be separated by rubbing one object with another. For macroscopic objects, negatively charged means an excess of electrons and positively charged means a depletion of electrons.
- The law of conservation of charge ensures that whenever a charge is created, an equal charge of the opposite sign is created at the same time.

Conceptual Questions

Exercise:

Problem:

There are very large numbers of charged particles in most objects. Why, then, don't most objects exhibit static electricity?

Exercise:

Problem:

Why do most objects tend to contain nearly equal numbers of positive and negative charges?

Problems & Exercises

Exercise:

Problem:

Common static electricity involves charges ranging from nanocoulombs to microcoulombs. (a) How many electrons are needed to form a charge of $-2.00\ \text{nC}$ (b) How many electrons must be removed from a neutral object to leave a net charge of $0.500\ \mu\text{C}$?

Solution:

(a) 1.25×10^{10}

(b) 3.13×10^{12}

Exercise:

Problem:

If 1.80×10^{20} electrons move through a pocket calculator during a full day's operation, how many coulombs of charge moved through it?

Exercise:

Problem:

To start a car engine, the car battery moves 3.75×10^{21} electrons through the starter motor. How many coulombs of charge were moved?

Solution:

-600 C

Exercise:

Problem:

A certain lightning bolt moves 40.0 C of charge. How many fundamental units of charge $\mid q_e \mid$ is this?

Glossary

electric charge
: a physical property of an object that causes it to be attracted toward or repelled from another charged object; each charged object generates and is influenced by a force called an electromagnetic force

law of conservation of charge
: states that whenever a charge is created, an equal amount of charge with the opposite sign is created simultaneously

electron

a particle orbiting the nucleus of an atom and carrying the smallest unit of negative charge

proton

a particle in the nucleus of an atom and carrying a positive charge equal in magnitude and opposite in sign to the amount of negative charge carried by an electron

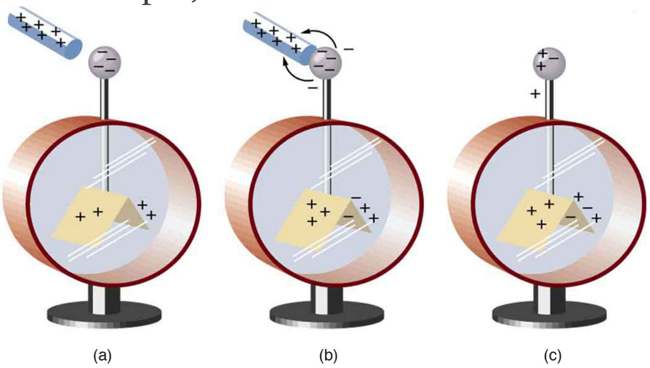
Conductors and Insulators

- Define conductor and insulator, explain the difference, and give examples of each.
- Describe three methods for charging an object.
- Explain what happens to an electric force as you move farther from the source.
- Define polarization.

This power adapter uses metal wires and connectors to conduct electricity from the wall socket to a laptop computer. The conducting wires allow electrons to move freely through the cables, which are shielded by rubber and plastic. These materials act as insulators that don’t allow electric charge to escape outward. (credit: Evan-Amos, Wikimedia Commons)

Some substances, such as metals and salty water, allow charges to move through them with relative ease. Some of the electrons in metals and similar conductors are not bound to individual atoms or sites in the material. These **free electrons** can move through the material much as air moves through loose sand. Any substance that has free electrons and allows charge to move relatively freely through it is called a **conductor**. The moving electrons may collide with fixed atoms and molecules, losing some energy, but they can move in a conductor. Superconductors allow the movement of charge without any loss of energy. Salty water and other similar conducting materials contain free ions that can move through them. An ion is an atom or molecule having a positive or negative (nonzero) total charge. In other words, the total number of electrons is not equal to the total number of protons.

Other substances, such as glass, do not allow charges to move through them. These are called **insulators**. Electrons and ions in insulators are bound in the structure and cannot move easily—as much as 10^{23} times more slowly than in conductors. Pure water and dry table salt are insulators, for example, whereas molten salt and salty water are conductors.



An electroscope is a favorite instrument in physics demonstrations and student laboratories. It is typically made with gold foil leaves hung from a (conducting) metal stem and is insulated from the room air in a glass-walled container. (a) A positively charged glass rod is brought near the tip of the electroscope, attracting electrons to the top and leaving a net positive charge on the leaves. Like charges in the light flexible gold leaves repel, separating them. (b) When the rod is touched against the ball, electrons are attracted and transferred, reducing the net charge on the glass rod but leaving the electroscope positively charged. (c) The excess charges are evenly distributed in the stem and leaves of the electroscope once the glass rod is removed.

Charging by Contact

[link] shows an electroscope being charged by touching it with a positively charged glass rod. Because the glass rod is an insulator, it must actually touch the electroscope to transfer charge to or from it. (Note that the extra positive charges reside on the surface of the glass rod as a result of rubbing it with silk before starting the experiment.) Since only electrons move in metals, we see that they are attracted to the top of the electroscope. There, some are transferred to the positive rod by touch, leaving the electroscope with a net positive charge.

Electrostatic repulsion in the leaves of the charged electroscope separates them. The electrostatic force has a horizontal component that results in the leaves moving apart as well as a vertical component that is balanced by the gravitational force. Similarly, the electroscope can be negatively charged by contact with a negatively charged object.

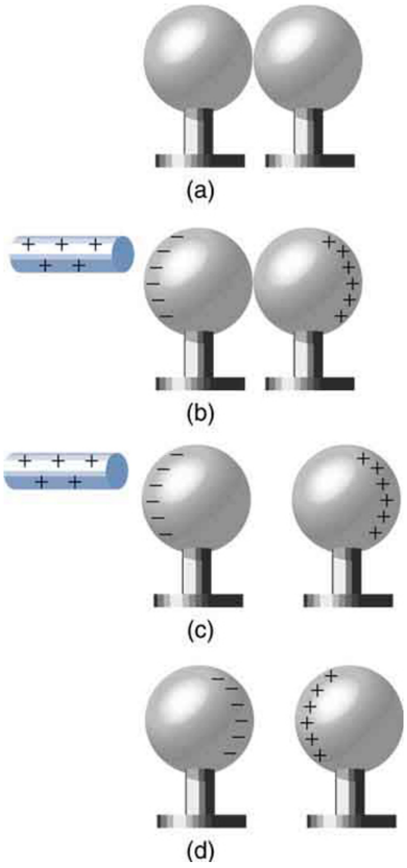
Charging by Induction

It is not necessary to transfer excess charge directly to an object in order to charge it. [link] shows a method of **induction** wherein a charge is created in a nearby object, without direct contact. Here we see two neutral metal spheres in contact with one another but insulated from the rest of the world.

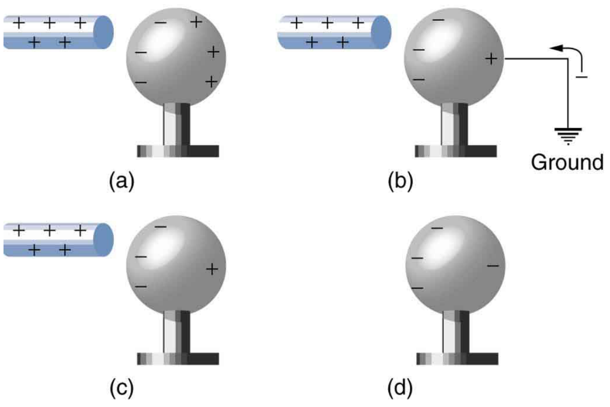
A positively charged rod is brought near one of them, attracting negative charge to that side, leaving the other sphere positively charged.

This is an example of induced **polarization** of neutral objects. Polarization is the separation of charges in an object that remains neutral. If the spheres are now separated (before the rod is pulled away), each sphere will have a net charge. Note that the object closest to the charged rod receives an opposite charge when charged by induction. Note also that no charge is removed from the charged rod, so that this process can be repeated without depleting the supply of excess charge.

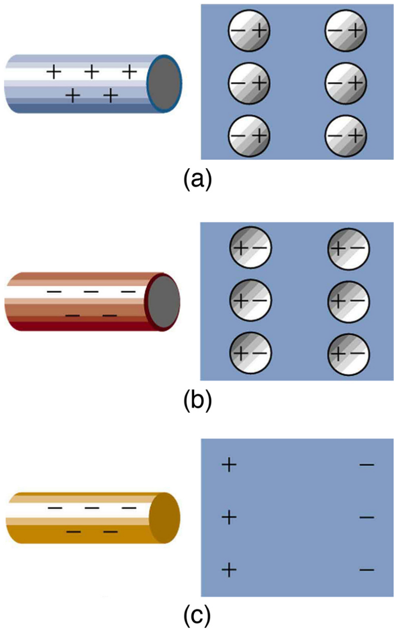
Another method of charging by induction is shown in [link]. The neutral metal sphere is polarized when a charged rod is brought near it. The sphere is then grounded, meaning that a conducting wire is run from the sphere to the ground. Since the earth is large and most ground is a good conductor, it can supply or accept excess charge easily. In this case, electrons are attracted to the sphere through a wire called the ground wire, because it supplies a conducting path to the ground. The ground connection is broken before the charged rod is removed, leaving the sphere with an excess charge opposite to that of the rod. Again, an opposite charge is achieved when charging by induction and the charged rod loses none of its excess charge.



Charging by induction. (a) Two uncharged or neutral metal spheres are in contact with each other but insulated from the rest of the world. (b) A positively charged glass rod is brought near the sphere on the left, attracting negative charge and leaving the other sphere positively charged. (c) The spheres are separated before the rod is removed, thus separating negative and positive charge. (d) The spheres retain net charges after the inducing rod is removed—without ever having been touched by a charged object.



Charging by induction, using a ground connection. (a) A positively charged rod is brought near a neutral metal sphere, polarizing it. (b) The sphere is grounded, allowing electrons to be attracted from the earth's ample supply. (c) The ground connection is broken. (d) The positive rod is removed, leaving the sphere with an induced negative charge.



Both positive and negative objects attract a neutral object by polarizing its molecules. (a) A positive object brought near a neutral insulator polarizes its molecules. There is a slight shift in the distribution of the electrons orbiting the molecule, with unlike charges being brought nearer and like charges moved away. Since the electrostatic force decreases with distance, there is a net attraction. (b) A negative object produces the opposite polarization, but again attracts the neutral object. (c) The same effect occurs for a conductor; since the unlike charges are closer, there is a net attraction.

Neutral objects can be attracted to any charged object. The pieces of straw attracted to polished amber are neutral, for example. If you run a plastic comb through your hair, the charged comb can pick up neutral pieces of paper. [link] shows how the polarization of atoms and molecules in neutral objects results in their attraction to a charged object.

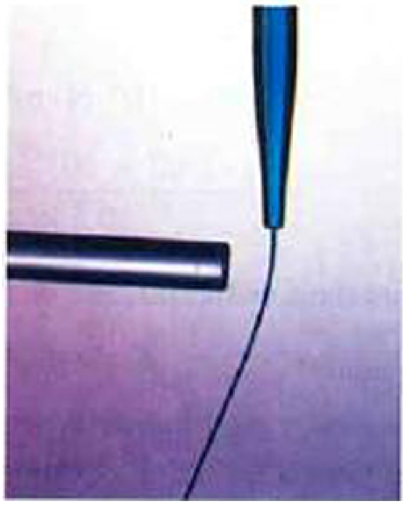
When a charged rod is brought near a neutral substance, an insulator in this case, the distribution of charge in atoms and molecules is shifted slightly. Opposite charge is attracted nearer the external charged rod, while like charge is repelled. Since the electrostatic force decreases with distance, the repulsion of like charges is weaker than the attraction of unlike charges, and so there is a net attraction. Thus a positively charged glass rod attracts neutral pieces of paper, as will a negatively charged rubber rod. Some molecules, like water, are polar molecules. Polar molecules have a natural or inherent separation of charge, although they are neutral overall. Polar molecules are particularly affected by other charged objects and show greater polarization effects than molecules with naturally uniform charge distributions.

Exercise:

Check Your Understanding

Problem:

Can you explain the attraction of water to the charged rod in the figure below?



Solution:

Answer

Water molecules are polarized, giving them slightly positive and slightly negative sides. This makes water even more susceptible to a charged rod’s attraction. As the water flows downward, due to the force of gravity, the charged conductor exerts a net attraction to the opposite charges in the stream of water, pulling it closer.

Note:

PhET Explorations: John Travoltage

Make sparks fly with John Travoltage. Wiggle Johnnie's foot and he picks up charges from the carpet. Bring his hand close to the door knob and get rid of the excess charge.
https://phet.colorado.edu/sims/html/john-travoltage/latest/john-travoltage_en.html

Section Summary

- Polarization is the separation of positive and negative charges in a neutral object.
- A conductor is a substance that allows charge to flow freely through its atomic structure.
- An insulator holds charge within its atomic structure.
- Objects with like charges repel each other, while those with unlike charges attract each other.
- A conducting object is said to be grounded if it is connected to the Earth through a conductor. Grounding allows transfer of charge to and from the earth’s large reservoir.
- Objects can be charged by contact with another charged object and obtain the same sign charge.
- If an object is temporarily grounded, it can be charged by induction, and obtains the opposite sign charge.
- Polarized objects have their positive and negative charges concentrated in different areas, giving them a non-symmetrical charge.
- Polar molecules have an inherent separation of charge.

Conceptual Questions

Exercise:

Problem:

An eccentric inventor attempts to levitate by first placing a large negative charge on himself and then putting a large positive charge on the ceiling of his workshop. Instead, while attempting to place a large negative charge on himself, his clothes fly off. Explain.

Exercise:

Problem:

If you have charged an electroscope by contact with a positively charged object, describe how you could use it to determine the charge of other objects. Specifically, what would the leaves of the electroscope do if other charged objects were brought near its knob?

Exercise:

Problem:

When a glass rod is rubbed with silk, it becomes positive and the silk becomes negative—yet both attract dust. Does the dust have a third type of charge that is attracted to both positive and negative? Explain.

Exercise:

Problem:

Why does a car always attract dust right after it is polished? (Note that car wax and car tires are insulators.)

Exercise:

Problem:

Describe how a positively charged object can be used to give another object a negative charge. What is the name of this process?

Exercise:

Problem:

What is grounding? What effect does it have on a charged conductor? On a charged insulator?

Problems & Exercises

Exercise:

Problem:

Suppose a speck of dust in an electrostatic precipitator has 1.0000×10^{12} protons in it and has a net charge of –5.00 nC (a very large charge for a small speck). How many electrons does it have?

Solution:

1.03×10^{12}

Exercise:

Problem:

An amoeba has 1.00×10^{16} protons and a net charge of 0.300 pC. (a) How many fewer electrons are there than protons? (b) If you paired them up, what fraction of the protons would have no electrons?

Exercise:

Problem:

A 50.0 g ball of copper has a net charge of $2.00\ \mu\text{C}$. What fraction of the copper's electrons has been removed? (Each copper atom has 29 protons, and copper has an atomic mass of 63.5.)

Solution:

9.09×10^{-13}

Exercise:

Problem:

What net charge would you place on a 100 g piece of sulfur if you put an extra electron on 1 in 10^{12} of its atoms? (Sulfur has an atomic mass of 32.1.)

Exercise:

Problem:

How many coulombs of positive charge are there in 4.00 kg of plutonium, given its atomic mass is 244 and that each plutonium atom has 94 protons?

Solution:

$1.48 \times 10^8 \text{ C}$

Glossary

free electron
: an electron that is free to move away from its atomic orbit

conductor
: a material that allows electrons to move separately from their atomic orbits

insulator
: a material that holds electrons securely within their atomic orbits

grounded
: when a conductor is connected to the Earth, allowing charge to freely flow to and from Earth’s unlimited reservoir

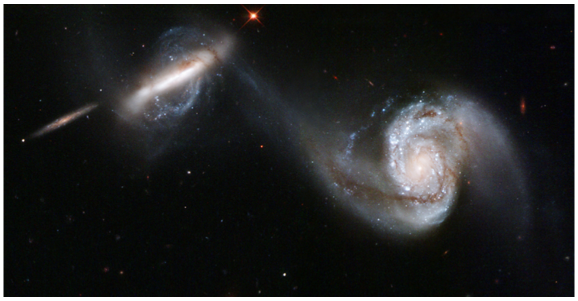
induction
: the process by which an electrically charged object brought near a neutral object creates a charge in that object

polarization
: slight shifting of positive and negative charges to opposite sides of an atom or molecule

electrostatic repulsion
: the phenomenon of two objects with like charges repelling each other

Coulomb’s Law

- State Coulomb’s law in terms of how the electrostatic force changes with the distance between two objects.
- Calculate the electrostatic force between two charged point forces, such as electrons or protons.
- Compare the electrostatic force to the gravitational attraction for a proton and an electron; for a human and the Earth.



This NASA image of Arp 87 shows the result of a strong gravitational attraction between two galaxies. In contrast, at the subatomic level, the electrostatic attraction between two objects, such as an electron and a proton, is far greater than their mutual attraction due to gravity. (credit: NASA/HST)

Through the work of scientists in the late 18th century, the main features of the **electrostatic force**—the existence of two types of charge, the observation that like charges repel, unlike charges attract, and the decrease of force with distance—were eventually refined, and expressed as a mathematical formula. The mathematical formula for the electrostatic force is called **Coulomb’s law** after the French physicist Charles Coulomb (1736–1806), who performed experiments and first proposed a formula to calculate it.

Note:
Coulomb’s Law
Equation:

$$F=k\frac{|q_1q_2|}{r^2}.$$

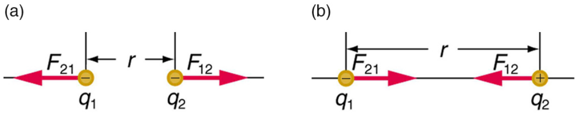
Coulomb’s law calculates the magnitude of the force F between two point charges, q_1 and q_2, separated by a distance r. In SI units, the constant k is equal to

Equation:

$$k = 8.988 \times 10^9 \frac{\mathrm{N} \cdot \mathrm{m}^2}{\mathrm{C}^2} \approx 8.99 \times 10^9 \frac{\mathrm{N} \cdot \mathrm{m}^2}{\mathrm{C}^2}.$$

The electrostatic force is a vector quantity and is expressed in units of newtons. The force is understood to be along the line joining the two charges. (See [link].)

Although the formula for Coulomb’s law is simple, it was no mean task to prove it. The experiments Coulomb did, with the primitive equipment then available, were difficult. Modern experiments have verified Coulomb’s law to great precision. For example, it has been shown that the force is inversely proportional to distance between two objects squared $\left(F \propto 1/r^2\right)$ to an accuracy of 1 part in 10^{16}. No exceptions have ever been found, even at the small distances within the atom.



The magnitude of the electrostatic force F between point charges q_1 and q_2 separated by a distance r is given by Coulomb’s law. Note that Newton’s third law (every force exerted creates an equal and opposite force) applies as usual—the force on q_1 is equal in magnitude and opposite in direction to the force it exerts on q_2. (a) Like charges. (b) Unlike charges.

Example:

How Strong is the Coulomb Force Relative to the Gravitational Force?

Compare the electrostatic force between an electron and proton separated by $0.530 \times 10^{-10}\ \mathrm{m}$ with the gravitational force between them. This distance is their average separation in a hydrogen atom.

Strategy

To compare the two forces, we first compute the electrostatic force using Coulomb’s law, $F = k\frac{|q_1 q_2|}{r^2}$. We then calculate the gravitational force using Newton’s universal law of gravitation. Finally, we take a ratio to see how the forces compare in magnitude.

Solution

Entering the given and known information about the charges and separation of the electron and proton into the expression of Coulomb’s law yields

Equation:

$$F = k\frac{|q_1 q_2|}{r^2}$$

Equation:

$$= \left(8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2\right) \times \frac{(1.60 \times 10^{-19}\text{ C})(1.60 \times 10^{-19}\text{ C})}{(0.530 \times 10^{-10}\text{ m})^2}$$

Thus the Coulomb force is

Equation:

$$F = 8.19 \times 10^{-8}\text{ N}.$$

The charges are opposite in sign, so this is an attractive force. This is a very large force for an electron—it would cause an acceleration of $8.99 \times 10^{22}\text{ m/s}^2$(verification is left as an end-of-section problem).The gravitational force is given by Newton’s law of gravitation as:

Equation:

$$F_G = G\frac{mM}{r^2},$$

where $G = 6.67 \times 10^{-11}\text{ N} \cdot \text{m}^2/\text{kg}^2$. Here m and M represent the electron and proton masses, which can be found in the appendices. Entering values for the knowns yields

Equation:

$$F_G = (6.67 \times 10^{-11}\text{ N} \cdot \text{m}^2/\text{kg}^2) \times \frac{(9.11 \times 10^{-31}\text{ kg})(1.67 \times 10^{-27}\text{ kg})}{(0.530 \times 10^{-10}\text{ m})^2} = 3.61 \times 10^{-47}\text{ N}$$

This is also an attractive force, although it is traditionally shown as positive since gravitational force is always attractive. The ratio of the magnitude of the electrostatic force to gravitational force in this case is, thus,

Equation:

$$\frac{F}{F_G} = 2.27 \times 10^{39}.$$

Discussion

This is a remarkably large ratio! Note that this will be the ratio of electrostatic force to gravitational force for an electron and a proton at any distance (taking the ratio before entering numerical values shows that the distance cancels). This ratio gives some indication of just how much larger the Coulomb force is than the gravitational force between two of the most common particles in nature.

As the example implies, gravitational force is completely negligible on a small scale, where the interactions of individual charged particles are important. On a large scale, such as between the Earth and a person, the reverse is true. Most objects are nearly electrically neutral, and so attractive and repulsive **Coulomb forces** nearly cancel. Gravitational force on a large scale dominates interactions between large objects because it is always attractive, while Coulomb forces tend to cancel.

Section Summary

- Frenchman Charles Coulomb was the first to publish the mathematical equation that describes the electrostatic force between two objects.
- Coulomb’s law gives the magnitude of the force between point charges. It is **Equation:**

  $$F = k\frac{|q_1 q_2|}{r^2},$$

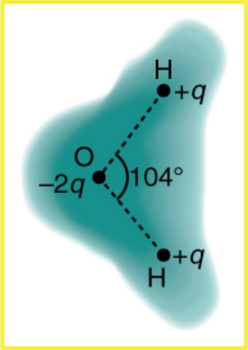
  where q_1 and q_2 are two point charges separated by a distance r, and $k \approx 8.99 \times 10^9 \text{ N} \cdot \text{m}^2/\text{C}^2$
- This Coulomb force is extremely basic, since most charges are due to point-like particles. It is responsible for all electrostatic effects and underlies most macroscopic forces.
- The Coulomb force is extraordinarily strong compared with the gravitational force, another basic force—but unlike gravitational force it can cancel, since it can be either attractive or repulsive.
- The electrostatic force between two subatomic particles is far greater than the gravitational force between the same two particles.

Conceptual Questions

Exercise:

Problem:

[link] shows the charge distribution in a water molecule, which is called a polar molecule because it has an inherent separation of charge. Given water’s polar character, explain what effect humidity has on removing excess charge from objects.



Schematic representation of the outer electron cloud of a neutral water molecule. The electrons spend more time near the oxygen than the hydrogens, giving a permanent charge separation as shown. Water is thus a *polar molecule*. It is more easily affected by electrostatic forces than molecules with uniform charge distributions.

Exercise:

Problem:

Using [link], explain, in terms of Coulomb’s law, why a polar molecule (such as in [link]) is attracted by both positive and negative charges.

Exercise:

Problem:

Given the polar character of water molecules, explain how ions in the air form nucleation centers for rain droplets.

Problems & Exercises

Exercise:

Problem:

What is the repulsive force between two pith balls that are 8.00 cm apart and have equal charges of – 30.0 nC?

Exercise:

Problem:

(a) How strong is the attractive force between a glass rod with a $0.700\ \mu\text{C}$ charge and a silk cloth with a –$0.600\ \mu\text{C}$ charge, which are 12.0 cm apart, using the approximation that they act like point charges? (b) Discuss how the answer to this problem might be affected if the charges are distributed over some area and do not act like point charges.

Solution:

(a) 0.263 N

(b) If the charges are distributed over some area, there will be a concentration of charge along the side closest to the oppositely charged object. This effect will increase the net force.

Exercise:

Problem:

Two point charges exert a 5.00 N force on each other. What will the force become if the distance between them is increased by a factor of three?

Exercise:

Problem:

Two point charges are brought closer together, increasing the force between them by a factor of 25. By what factor was their separation decreased?

Solution:

The separation decreased by a factor of 5.

Exercise:

Problem:

How far apart must two point charges of 75.0 nC (typical of static electricity) be to have a force of 1.00 N between them?

Exercise:

Problem:

If two equal charges each of 1 C each are separated in air by a distance of 1 km, what is the magnitude of the force acting between them? You will see that even at a distance as large as 1 km, the repulsive force is substantial because 1 C is a very significant amount of charge.

Exercise:

Problem:

A test charge of $+2\ \mu\text{C}$ is placed halfway between a charge of $+6\ \mu\text{C}$ and another of $+4\ \mu\text{C}$ separated by 10 cm. (a) What is the magnitude of the force on the test charge? (b) What is the direction of this force (away from or toward the $+6\ \mu\text{C}$ charge)?

Exercise:

Problem:

Bare free charges do not remain stationary when close together. To illustrate this, calculate the acceleration of two isolated protons separated by 2.00 nm (a typical distance between gas atoms). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.

Solution:

$$\begin{aligned} F &= k\frac{|q_1q_2|}{r^2} = ma \Rightarrow a = \frac{kq^2}{mr^2} \\ &= \frac{\left(9.00\times10^9\ \text{N}\cdot\text{m}^2/\text{C}^2\right)\left(1.60\times10^{-19}\ \text{m}\right)^2}{\left(1.67\times10^{-27}\ \text{kg}\right)\left(2.00\times10^{-9}\ \text{m}\right)^2} \\ &= 3.45\times10^{16}\ \text{m/s}^2 \end{aligned}$$

Exercise:

Problem:

(a) By what factor must you change the distance between two point charges to change the force between them by a factor of 10? (b) Explain how the distance can either increase or decrease by this factor and still cause a factor of 10 change in the force.

Solution:

(a) 3.2

(b) If the distance increases by 3.2, then the force will decrease by a factor of 10 ; if the distance decreases by 3.2, then the force will increase by a factor of 10. Either way, the force changes by a factor of 10.

Exercise:

Problem:

Suppose you have a total charge q_{tot} that you can split in any manner. Once split, the separation distance is fixed. How do you split the charge to achieve the greatest force?

Exercise:

Problem:

(a) Common transparent tape becomes charged when pulled from a dispenser. If one piece is placed above another, the repulsive force can be great enough to support the top piece's weight. Assuming equal point charges (only an approximation), calculate the magnitude of the charge if electrostatic force is great enough to support the weight of a 10.0 mg piece of tape held 1.00 cm above another. (b) Discuss whether the magnitude of this charge is consistent with what is typical of static electricity.

Solution:

(a) 1.04×10^{-9} C

(b) This charge is approximately 1 nC, which is consistent with the magnitude of charge typical for static electricity

Exercise:

Problem:

(a) Find the ratio of the electrostatic to gravitational force between two electrons. (b) What is this ratio for two protons? (c) Why is the ratio different for electrons and protons?

Exercise:

Problem:

At what distance is the electrostatic force between two protons equal to the weight of one proton?

Exercise:

Problem:

A certain five cent coin contains 5.00 g of nickel. What fraction of the nickel atoms' electrons, removed and placed 1.00 m above it, would support the weight of this coin? The atomic mass of nickel is 58.7, and each nickel atom contains 28 electrons and 28 protons.

Solution:

1.02×10^{-11}

Exercise:

Problem:

(a) Two point charges totaling $8.00\ \mu\text{C}$ exert a repulsive force of 0.150 N on one another when separated by 0.500 m. What is the charge on each? (b) What is the charge on each if the force is attractive?

Exercise:

Problem:

Point charges of $5.00\ \mu\text{C}$ and $-3.00\ \mu\text{C}$ are placed 0.250 m apart. (a) Where can a third charge be placed so that the net force on it is zero? (b) What if both charges are positive?

Solution:

a. 0.859 m beyond negative charge on line connecting two charges
b. 0.109 m from lesser charge on line connecting two charges

Exercise:

Problem:

Two point charges q_1 and q_2 are $3.00\ \text{m}$ apart, and their total charge is $20\ \mu\text{C}$. (a) If the force of repulsion between them is 0.075N, what are magnitudes of the two charges? (b) If one charge attracts the other with a force of 0.525N, what are the magnitudes of the two charges? Note that you may need to solve a quadratic equation to reach your answer.

Glossary

Coulomb's law
: the mathematical equation calculating the electrostatic force vector between two charged particles

Coulomb force
: another term for the electrostatic force

electrostatic force

the amount and direction of attraction or repulsion between two charged bodies

Electric Field: Concept of a Field Revisited

- Describe a force field and calculate the strength of an electric field due to a point charge.
- Calculate the force exerted on a test charge by an electric field.
- Explain the relationship between electrical force (F) on a test charge and electrical field strength (E).

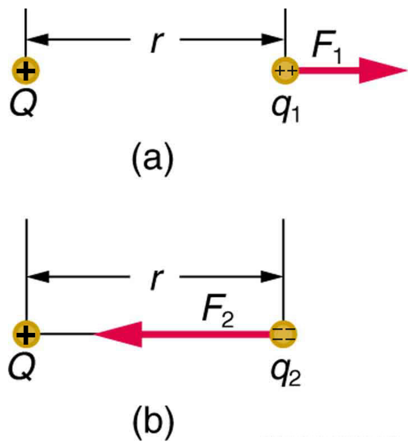
Contact forces, such as between a baseball and a bat, are explained on the small scale by the interaction of the charges in atoms and molecules in close proximity. They interact through forces that include the **Coulomb force**. Action at a distance is a force between objects that are not close enough for their atoms to “touch.” That is, they are separated by more than a few atomic diameters.

For example, a charged rubber comb attracts neutral bits of paper from a distance via the Coulomb force. It is very useful to think of an object being surrounded in space by a **force field**. The force field carries the force to another object (called a test object) some distance away.

Concept of a Field

A field is a way of conceptualizing and mapping the force that surrounds any object and acts on another object at a distance without apparent physical connection. For example, the gravitational field surrounding the earth (and all other masses) represents the gravitational force that would be experienced if another mass were placed at a given point within the field.

In the same way, the Coulomb force field surrounding any charge extends throughout space. Using Coulomb’s law, $F=k|q_1q_2|/r^2$, its magnitude is given by the equation $F=k|\ qQ\ |/r^2$, for a **point charge** (a particle having a charge Q) acting on a **test charge** q at a distance r (see [link]). Both the magnitude and direction of the Coulomb force field depend on Q and the test charge q.



The Coulomb force field due to a positive charge Q is shown acting on two different charges. Both charges are the same distance from Q. (a) Since q_1 is positive, the force F_1 acting on it is repulsive. (b) The charge q_2 is negative and greater in magnitude than q_1, and so the force F_2 acting on it is attractive and stronger than F_1. The Coulomb force field is thus not unique at any point in space, because it depends on the test charges q_1 and q_2

as well as the
charge Q.

To simplify things, we would prefer to have a field that depends only on Q and not on the test charge q. The electric field is defined in such a manner that it represents only the charge creating it and is unique at every point in space. Specifically, the electric field E is defined to be the ratio of the Coulomb force to the test charge:

Equation:

$$\mathbf{E} = \frac{\mathbf{F}}{q},$$

where $\mathbf{F}$ is the electrostatic force (or Coulomb force) exerted on a positive test charge q. It is understood that $\mathbf{E}$ is in the same direction as $\mathbf{F}$. It is also assumed that q is so small that it does not alter the charge distribution creating the electric field. The units of electric field are newtons per coulomb (N/C). If the electric field is known, then the electrostatic force on any charge q is simply obtained by multiplying charge times electric field, or $\mathbf{F} = q\mathbf{E}$. Consider the electric field due to a point charge Q. According to Coulomb's law, the force it exerts on a test charge q is $F = k| \ qQ \ |/r^2$. Thus the magnitude of the electric field, E, for a point charge is

Equation:

$$E = \left|\frac{F}{q}\right| = k\left|\frac{\text{qQ}}{qr^2}\right| = k\frac{|Q|}{r^2}.$$

Since the test charge cancels, we see that

Equation:

$$E = k\frac{|Q|}{r^2}.$$

The electric field is thus seen to depend only on the charge Q and the distance r; it is completely independent of the test charge q.

Example:
Calculating the Electric Field of a Point Charge

Calculate the strength and direction of the electric field E due to a point charge of 2.00 nC (nano-Coulombs) at a distance of 5.00 mm from the charge.

Strategy

We can find the electric field created by a point charge by using the equation $E = \mathrm{kQ}/r^2$.

Solution

Here $Q = 2.00 \times 10^{-9}$ C and $r = 5.00 \times 10^{-3}$ m. Entering those values into the above equation gives

Equation:

$$\begin{aligned} E &= k\frac{Q}{r^2} \\ &= (8.99 \times 10^9\ \mathrm{N \cdot m^2/C^2}) \times \frac{(2.00 \times 10^{-9}\ \mathrm{C})}{(5.00 \times 10^{-3}\ \mathrm{m})^2} \\ &= 7.19 \times 10^5\ \mathrm{N/C}. \end{aligned}$$

Discussion

This **electric field strength** is the same at any point 5.00 mm away from the charge Q that creates the field. It is positive, meaning that it has a direction pointing away from the charge Q.

Example:
Calculating the Force Exerted on a Point Charge by an Electric Field

What force does the electric field found in the previous example exert on a point charge of –0.250 μC?

Strategy

Since we know the electric field strength and the charge in the field, the force on that charge can be calculated using the definition of electric field

$\mathbf{E} = \mathbf{F}/q$ rearranged to $\mathbf{F} = q\mathbf{E}$.

Solution

The magnitude of the force on a charge $q = -0.250\ \mu\text{C}$ exerted by a field of strength $E = 7.20 \times 10^5$ N/C is thus,

Equation:

$$\begin{aligned} F &= -qE \\ &= (0.250 \times 10^{-6}\ \text{C})(7.20 \times 10^{5}\ \text{N/C}) \\ &= 0.180\ \text{N}. \end{aligned}$$

Because q is negative, the force is directed opposite to the direction of the field.

Discussion

The force is attractive, as expected for unlike charges. (The field was created by a positive charge and here acts on a negative charge.) The charges in this example are typical of common static electricity, and the modest attractive force obtained is similar to forces experienced in static cling and similar situations.

Note:

PhET Explorations: Electric Field of Dreams

Play ball! Add charges to the Field of Dreams and see how they react to the electric field. Turn on a background electric field and adjust the direction and magnitude.

https://archive.cnx.org/specials/ca9a78b4-06a7-11e6-b638-3bb71d1f0b42/electric-field-of-dreams/#sim-electric-field-of-dreams

Section Summary

- The electrostatic force field surrounding a charged object extends out into space in all directions.
- The electrostatic force exerted by a point charge on a test charge at a distance r depends on the charge of both charges, as well as the

distance between the two.

- The electric field $\mathbf{E}$ is defined to be

  Equation:

  $$\mathbf{E} = \frac{\mathbf{F}}{q,}$$

  where $\mathbf{F}$ is the Coulomb or electrostatic force exerted on a small positive test charge q. $\mathbf{E}$ has units of N/C.
- The magnitude of the electric field $\mathbf{E}$ created by a point charge Q is

  Equation:

  $$\mathbf{E} = k\frac{|Q|}{r^2}.$$

  where r is the distance from Q. The electric field $\mathbf{E}$ is a vector and fields due to multiple charges add like vectors.

Conceptual Questions

Exercise:

Problem:

Why must the test charge q in the definition of the electric field be vanishingly small?

Exercise:

Problem:

Are the direction and magnitude of the Coulomb force unique at a given point in space? What about the electric field?

Problem Exercises

Exercise:

Problem:

What is the magnitude and direction of an electric field that exerts a 2.00×10^{-5} N upward force on a –$1.75\ \mu\text{C}$ charge?

Exercise:

Problem:

What is the magnitude and direction of the force exerted on a $3.50\ \mu\text{C}$ charge by a 250 N/C electric field that points due east?

Solution:

8.75×10^{-4} N

Exercise:

Problem:

Calculate the magnitude of the electric field 2.00 m from a point charge of 5.00 mC (such as found on the terminal of a Van de Graaff).

Exercise:

Problem:

(a) What magnitude point charge creates a 10,000 N/C electric field at a distance of 0.250 m? (b) How large is the field at 10.0 m?

Solution:

(a) 6.94×10^{-8} C

(b) $6.25\ \text{N}/\text{C}$

Exercise:

Problem:

Calculate the initial (from rest) acceleration of a proton in a $5.00 \times 10^{6}\ \text{N/C}$ electric field (such as created by a research Van de Graaff). Explicitly show how you follow the steps in the Problem-Solving Strategy for electrostatics.

Exercise:

Problem:

(a) Find the magnitude and direction of an electric field that exerts a $4.80 \times 10^{-17}\ \text{N}$ westward force on an electron. (b) What magnitude and direction force does this field exert on a proton?

Solution:

(a) $300\ \text{N/C (east)}$

(b) $4.80 \times 10^{-17}\ \text{N (east)}$

Glossary

field
: a map of the amount and direction of a force acting on other objects, extending out into space

point charge
: A charged particle, designated Q, generating an electric field

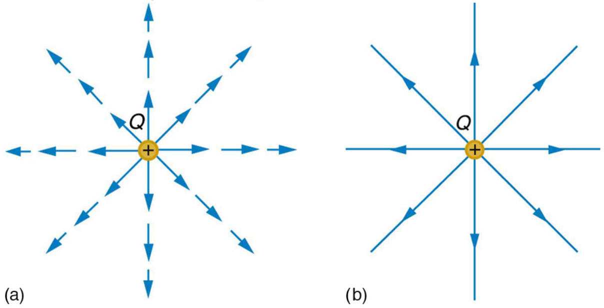
test charge
: A particle (designated q) with either a positive or negative charge set down within an electric field generated by a point charge

Electric Field Lines: Multiple Charges

- Calculate the total force (magnitude and direction) exerted on a test charge from more than one charge
- Describe an electric field diagram of a positive point charge; of a negative point charge with twice the magnitude of positive charge
- Draw the electric field lines between two points of the same charge; between two points of opposite charge.

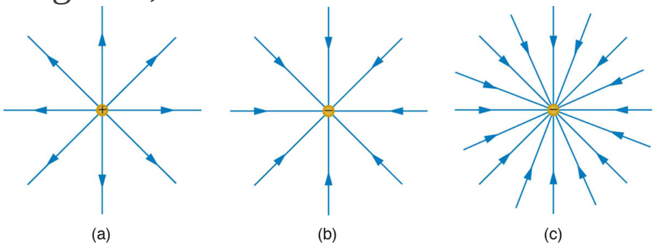
Drawings using lines to represent **electric fields** around charged objects are very useful in visualizing field strength and direction. Since the electric field has both magnitude and direction, it is a vector. Like all **vectors**, the electric field can be represented by an arrow that has length proportional to its magnitude and that points in the correct direction. (We have used arrows extensively to represent force vectors, for example.)

[link] shows two pictorial representations of the same electric field created by a positive point charge Q. [link] (b) shows the standard representation using continuous lines. [link] (a) shows numerous individual arrows with each arrow representing the force on a test charge q. Field lines are essentially a map of infinitesimal force vectors.



Two equivalent representations of the electric field due to a positive charge Q. (a) Arrows representing the electric field's magnitude and direction. (b) In the standard representation, the arrows are replaced by continuous field lines having the same direction at any point as the electric field. The closeness of the lines is directly related to the strength of the electric field. A test charge placed anywhere will feel a force in the direction of the field line; this force will have a strength proportional to the density of the lines (being greater near the charge, for example).

Note that the electric field is defined for a positive test charge q, so that the field lines point away from a positive charge and toward a negative charge. (See [link].) The electric field strength is exactly proportional to the number of field lines per unit area, since the magnitude of the electric field for a point charge is $E = k|Q|/r^2$ and area is proportional to r^2. This pictorial representation, in which field lines represent the direction and their closeness (that is, their areal density or the number of lines crossing a unit area) represents strength, is used for all fields: electrostatic, gravitational, magnetic, and others.



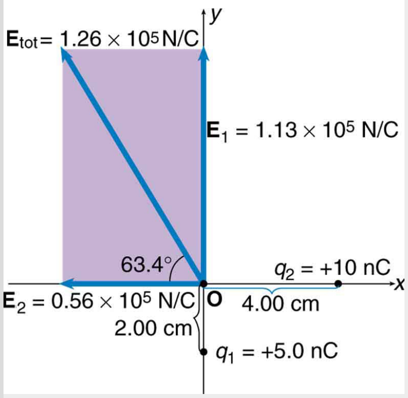
The electric field surrounding three different point charges. (a) A positive charge. (b) A negative charge of equal magnitude. (c) A larger negative charge.

In many situations, there are multiple charges. The total electric field created by multiple charges is the vector sum of the individual fields created by each charge. The following example shows how to add electric field vectors.

Example:

Adding Electric Fields

Find the magnitude and direction of the total electric field due to the two point charges, q_1 and q_2, at the origin of the coordinate system as shown in [link].



y
$\mathbf{E}_{\text{tot}} = 1.26 \times 10^5$ N/C
$\mathbf{E}_1 = 1.13 \times 10^5$ N/C
63.4°
$q_2 = +10$ nC
x
$\mathbf{E}_2 = 0.56 \times 10^5$ N/C
O
4.00 cm
2.00 cm
$q_1 = +5.0$ nC

The electric fields $\mathbf{E}_1$ and $\mathbf{E}_2$ at the origin O add to $\mathbf{E}_{\text{tot}}$.

Strategy

Since the electric field is a vector (having magnitude and direction), we add electric fields with the same vector techniques used for other types of vectors. We first must find the electric field due to each charge at the point of interest, which is the origin of the coordinate system (O) in this instance. We pretend that there is a positive test charge, q, at point O, which allows us to determine the direction of the fields $\mathbf{E}_1$ and $\mathbf{E}_2$. Once those fields are found, the total field can be determined using **vector addition**.

Solution

The electric field strength at the origin due to q_1 is labeled E_1 and is calculated:

Equation:

$$E_1 = k\frac{q_1}{r_1^2} = \left(8.99 \times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2\right)\frac{(5.00\times10^{-9}\text{ C})}{(2.00\times10^{-2}\text{ m})^2}$$
$$E_1 = 1.124 \times 10^5 \text{ N/C}.$$

Similarly, E_2 is

Equation:

$$E_2 = k\frac{q_2}{r_2^2} = \left(8.99 \times 10^9 \text{ N}\cdot\text{m}^2/\text{C}^2\right)\frac{(10.0\times10^{-9}\text{ C})}{(4.00\times10^{-2}\text{ m})^2}$$
$$E_2 = 0.5619 \times 10^5 \text{ N/C}.$$

Four digits have been retained in this solution to illustrate that E_1 is exactly twice the magnitude of E_2. Now arrows are drawn to represent the magnitudes and directions of $\mathbf{E}_1$ and $\mathbf{E}_2$. (See [link].) The direction of the electric field is that of the force on a positive charge so both arrows point directly away from the positive charges that create them. The arrow for $\mathbf{E}_1$ is exactly twice the length of that for $\mathbf{E}_2$. The arrows form a right triangle in this case and can be added using the Pythagorean theorem. The magnitude of the total field E_{tot} is

Equation:

$$\begin{aligned} E_{\text{tot}} &= (E_1^2 + E_2^2)^{1/2} \\ &= \{(1.124 \times 10^5 \text{ N/C})^2 + (0.5619 \times 10^5 \text{ N/C})^2\}^{1/2} \\ &= 1.26 \times 10^5 \text{ N/C}. \end{aligned}$$

The direction is

Equation:

$$\begin{aligned}\theta &= \tan^{-1}\left(\frac{E_1}{E_2}\right)\\ &= \tan^{-1}\left(\frac{1.124\times10^5\ \text{N/C}}{0.5619\times10^5\ \text{N/C}}\right)\\ &= 63.4^{\text{o}},\end{aligned}$$

or 63.4^{o} above the *x*-axis.

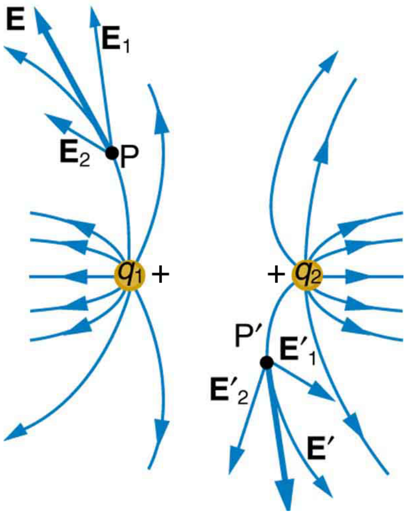
Discussion

In cases where the electric field vectors to be added are not perpendicular, vector components or graphical techniques can be used. The total electric field found in this example is the total electric field at only one point in space. To find the total electric field due to these two charges over an entire region, the same technique must be repeated for each point in the region. This impossibly lengthy task (there are an infinite number of points in space) can be avoided by calculating the total field at representative points and using some of the unifying features noted next.

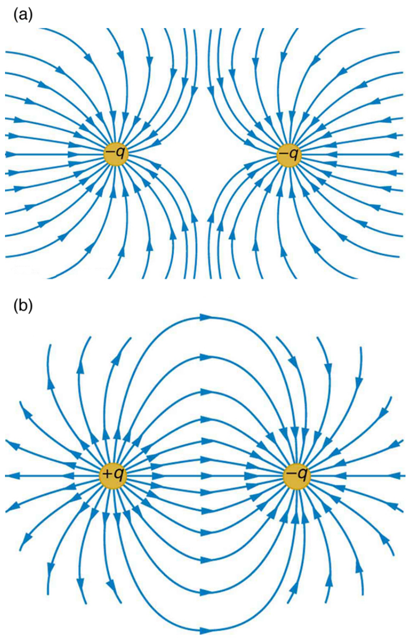
[link] shows how the electric field from two point charges can be drawn by finding the total field at representative points and drawing electric field lines consistent with those points. While the electric fields from multiple charges are more complex than those of single charges, some simple features are easily noticed.

For example, the field is weaker between like charges, as shown by the lines being farther apart in that region. (This is because the fields from each charge exert opposing forces on any charge placed between them.) (See [link] and [link](a).) Furthermore, at a great distance from two like charges, the field becomes identical to the field from a single, larger charge.

[link](b) shows the electric field of two unlike charges. The field is stronger between the charges. In that region, the fields from each charge are in the same direction, and so their strengths add. The field of two unlike charges is weak at large distances, because the fields of the individual charges are in opposite directions and so their strengths subtract. At very large distances, the field of two unlike charges looks like that of a smaller single charge.



Two positive point charges q_1 and q_2 produce the resultant electric field shown. The field is calculated at representative points and then smooth field lines drawn following the rules outlined in the text.



(a) Two negative charges produce the fields shown. It is very similar to the field produced by two positive charges, except that the directions are reversed. The field is clearly weaker between the charges. The individual forces on a test charge in that region are in opposite directions. (b) Two opposite charges produce the field shown, which is stronger in the region between the charges.

We use electric field lines to visualize and analyze electric fields (the lines are a pictorial tool, not a physical entity in themselves). The properties of electric field lines for any charge distribution can be summarized as follows:

1. Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
2. The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
3. The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
4. The direction of the electric field is tangent to the field line at any point in space.
5. Field lines can never cross.

The last property means that the field is unique at any point. The field line represents the direction of the field; so if they crossed, the field would have two directions at that location (an impossibility if the field is unique).

Note:
PhET Explorations: Charges and Fields
Move point charges around on the playing field and then view the electric field, voltages, equipotential lines, and more. It's colorful, it's dynamic, it's free.

Click here for the simulation.

Section Summary

- Drawings of electric field lines are useful visual tools. The properties of electric field lines for any charge distribution are that:
- Field lines must begin on positive charges and terminate on negative charges, or at infinity in the hypothetical case of isolated charges.
- The number of field lines leaving a positive charge or entering a negative charge is proportional to the magnitude of the charge.
- The strength of the field is proportional to the closeness of the field lines—more precisely, it is proportional to the number of lines per unit area perpendicular to the lines.
- The direction of the electric field is tangent to the field line at any point in space.
- Field lines can never cross.

Conceptual Questions

Exercise:

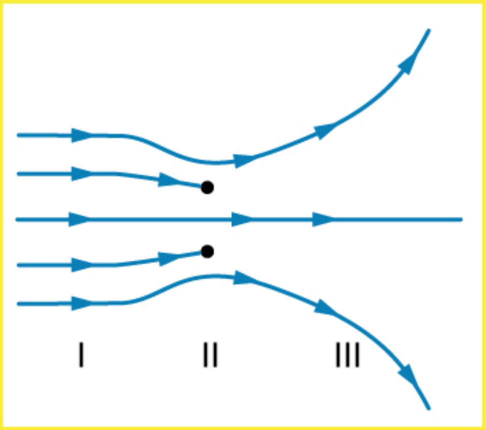
Problem:

Compare and contrast the Coulomb force field and the electric field. To do this, make a list of five properties for the Coulomb force field analogous to the five properties listed for electric field lines. Compare each item in your list of Coulomb force field properties with those of the electric field—are they the same or different? (For example, electric field lines cannot cross. Is the same true for Coulomb field lines?)

Exercise:

Problem:

[link] shows an electric field extending over three regions, labeled I, II, and III. Answer the following questions. (a) Are there any isolated charges? If so, in what region and what are their signs? (b) Where is the field strongest? (c) Where is it weakest? (d) Where is the field the most uniform?



Problem Exercises

Exercise:

Problem:

(a) Sketch the electric field lines near a point charge $+q$. (b) Do the same for a point charge $-3.00q$.

Exercise:

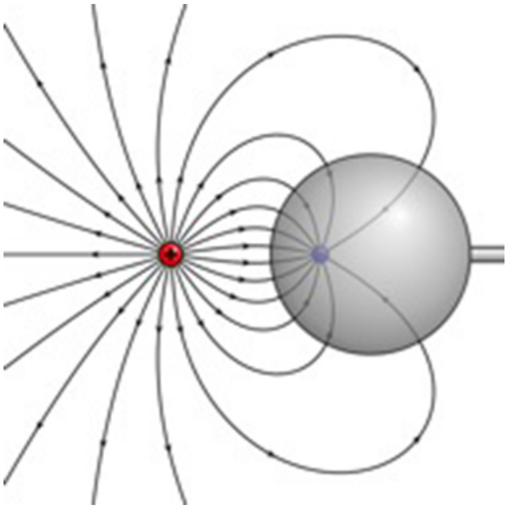
Problem:

Sketch the electric field lines a long distance from the charge distributions shown in [link] (a) and (b)

Exercise:

Problem:

[link] shows the electric field lines near two charges q_1 and q_2. What is the ratio of their magnitudes? (b) Sketch the electric field lines a long distance from the charges shown in the figure.



The electric field near two charges.

Exercise:

Problem:

Sketch the electric field lines in the vicinity of two opposite charges, where the negative charge is three times greater in magnitude than the positive. (See [link] for a similar situation).

Glossary

electric field
: a three-dimensional map of the electric force extended out into space from a point charge

electric field lines
: a series of lines drawn from a point charge representing the magnitude and direction of force exerted by that charge

vector
: a quantity with both magnitude and direction

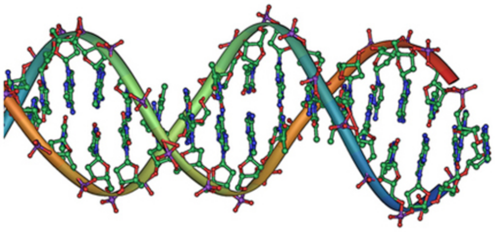
vector addition

mathematical combination of two or more vectors, including their magnitudes, directions, and positions

Electric Forces in Biology

- Describe how a water molecule is polar.
- Explain electrostatic screening by a water molecule within a living cell.

Classical electrostatics has an important role to play in modern molecular biology. Large molecules such as proteins, nucleic acids, and so on—so important to life—are usually electrically charged. DNA itself is highly charged; it is the electrostatic force that not only holds the molecule together but gives the molecule structure and strength. [link] is a schematic of the DNA double helix.



DNA is a highly charged molecule. The DNA double helix shows the two coiled strands each containing a row of nitrogenous bases, which “code” the genetic information needed by a living organism. The strands are connected by bonds between pairs of bases. While pairing combinations between certain bases are fixed (C-G and A-T), the sequence of nucleotides in the strand varies. (credit: Jerome Walker)

The four nucleotide bases are given the symbols A (adenine), C (cytosine), G (guanine), and T (thymine). The order of the four bases varies in each strand, but the pairing between bases is always the same. C and G are always paired and A and T are always paired, which helps to preserve the order of bases in cell division (mitosis) so as to pass on the correct genetic information. Since the Coulomb force drops with distance ($F \propto 1/r^2$), the distances between the base pairs must be small enough that the electrostatic force is sufficient to hold them together.

DNA is a highly charged molecule, with about $2q_{\text{e}}$ (fundamental charge) per 0.3×10^{-9} m. The distance separating the two strands that make up the DNA structure is about 1 nm, while the distance separating the individual atoms within each base is about 0.3 nm.

One might wonder why electrostatic forces do not play a larger role in biology than they do if we have so many charged molecules. The reason is that the electrostatic force is “diluted” due to **screening** between molecules. This is due to the presence of other charges in the cell.

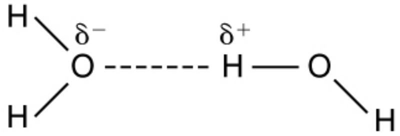
Polarity of Water Molecules

The best example of this charge screening is the water molecule, represented as H_2O. Water is a strongly **polar molecule**. Its 10 electrons (8 from the oxygen atom and 2 from the two hydrogen atoms) tend to remain closer to the oxygen nucleus than the hydrogen nuclei. This creates two centers of equal and opposite charges—what is called a **dipole**, as illustrated in [link]. The magnitude of the dipole is called the dipole moment.

These two centers of charge will terminate some of the electric field lines coming from a free charge, as on a DNA molecule. This results in a reduction in the strength of the **Coulomb interaction**. One might say that screening makes the Coulomb force a short range force rather than long range.

Other ions of importance in biology that can reduce or screen Coulomb interactions are Na^+, and K^+, and Cl^-. These ions are located both inside and outside of living cells. The movement of these ions through cell membranes is crucial to the motion of nerve impulses through nerve axons.

Recent studies of electrostatics in biology seem to show that electric fields in cells can be extended over larger distances, in spite of screening, by “microtubules” within the cell. These microtubules are hollow tubes composed of proteins that guide the movement of chromosomes when cells divide, the motion of other organisms within the cell, and provide mechanisms for motion of some cells (as motors).



This schematic shows water (H_2O) as a polar molecule. Unequal sharing of electrons between the oxygen (O) and hydrogen (H) atoms leads to a net separation of positive and negative charge—forming a dipole. The symbols δ^- and δ^+ indicate that the oxygen side of the H_2O molecule tends to be more negative, while the hydrogen ends tend to be more positive. This leads to an attraction of opposite charges between molecules.

Section Summary

- Many molecules in living organisms, such as DNA, carry a charge.
- An uneven distribution of the positive and negative charges within a polar molecule produces a dipole.
- The effect of a Coulomb field generated by a charged object may be reduced or blocked by other nearby charged objects.
- Biological systems contain water, and because water molecules are polar, they have a strong effect on other molecules in living systems.

Conceptual Question

Exercise:

Problem:

A cell membrane is a thin layer enveloping a cell. The thickness of the membrane is much less than the size of the cell. In a static situation the membrane has a charge distribution of $-2.5 \times 10^{-6}\text{C/m}^{\ 2}$ on its inner surface and $+2.5 \times 10^{-6}\ \text{C/m}^2$ on its outer surface. Draw a diagram of the cell and the surrounding cell membrane. Include on this diagram the charge distribution and the corresponding electric field. Is there any electric field inside the cell? Is there any electric field outside the cell?

Glossary

dipole

a molecule’s lack of symmetrical charge distribution, causing one side to be more positive and another to be more negative

polar molecule

a molecule with an asymmetrical distribution of positive and negative charge

screening

the dilution or blocking of an electrostatic force on a charged object by the presence of other charges nearby

Coulomb interaction

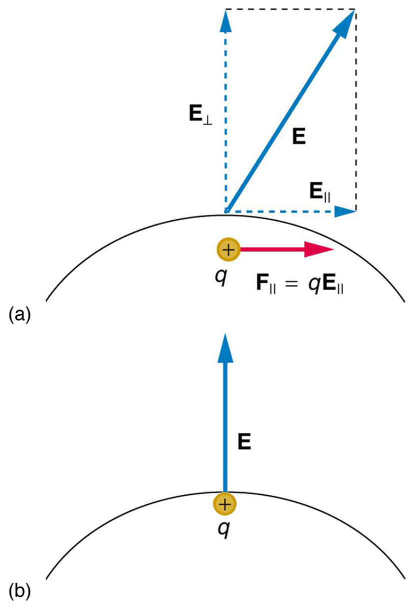
the interaction between two charged particles generated by the Coulomb forces they exert on one another

Conductors and Electric Fields in Static Equilibrium

- List the three properties of a conductor in electrostatic equilibrium.
- Explain the effect of an electric field on free charges in a conductor.
- Explain why no electric field may exist inside a conductor.
- Describe the electric field surrounding Earth.
- Explain what happens to an electric field applied to an irregular conductor.
- Describe how a lightning rod works.
- Explain how a metal car may protect passengers inside from the dangerous electric fields caused by a downed line touching the car.

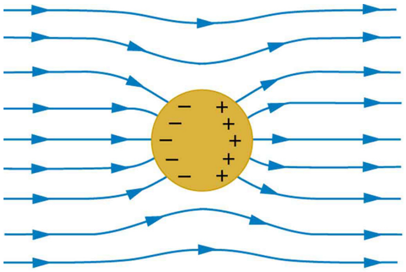
Conductors contain **free charges** that move easily. When excess charge is placed on a conductor or the conductor is put into a static electric field, charges in the conductor quickly respond to reach a steady state called **electrostatic equilibrium**.

[link] shows the effect of an electric field on free charges in a conductor. The free charges move until the field is perpendicular to the conductor's surface. There can be no component of the field parallel to the surface in electrostatic equilibrium, since, if there were, it would produce further movement of charge. A positive free charge is shown, but free charges can be either positive or negative and are, in fact, negative in metals. The motion of a positive charge is equivalent to the motion of a negative charge in the opposite direction.



When an electric field $\mathbf{E}$ is applied to a conductor, free charges inside the conductor move until the field is perpendicular to the surface. (a) The electric field is a vector quantity, with both parallel and perpendicular components. The parallel component $(\mathbf{E}_{||})$ exerts a force $(\mathbf{F}_{||})$ on the free charge q, which moves the charge until $\mathbf{F}_{||} = 0$. (b) The resulting field is perpendicular to the surface. The free charge has been brought to the conductor’s surface, leaving electrostatic forces in equilibrium.

A conductor placed in an **electric field** will be **polarized**. [link] shows the result of placing a neutral conductor in an originally uniform electric field. The field becomes stronger near the conductor but entirely disappears inside it.

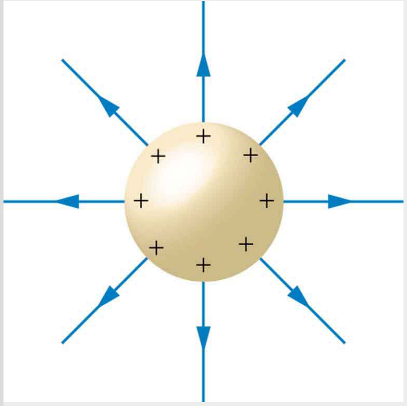


This illustration shows a spherical conductor in static equilibrium with an originally uniform electric field. Free charges move within the conductor, polarizing it, until the electric field lines are perpendicular to the surface. The field lines end on excess negative charge on one section of the surface and begin again on excess positive charge on the opposite side. No electric field exists inside the conductor, since free charges in the conductor would continue moving in response to any field until it was neutralized.

Note:

Misconception Alert: Electric Field inside a Conductor

Excess charges placed on a spherical conductor repel and move until they are evenly distributed, as shown in [link]. Excess charge is forced to the surface until the field inside the conductor is zero. Outside the conductor, the field is exactly the same as if the conductor were replaced by a point charge at its center equal to the excess charge.



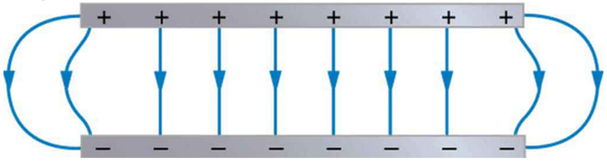
The mutual repulsion of excess positive charges on a spherical conductor distributes them uniformly on its surface. The resulting electric field is perpendicular to the surface and zero inside. Outside the conductor, the field is identical to that of a point charge at the center equal to the excess charge.

Note:
Properties of a Conductor in Electrostatic Equilibrium

1. The electric field is zero inside a conductor.
2. Just outside a conductor, the electric field lines are perpendicular to its surface, ending or beginning on charges on the surface.
3. Any excess charge resides entirely on the surface or surfaces of a conductor.

The properties of a conductor are consistent with the situations already discussed and can be used to analyze any conductor in electrostatic equilibrium. This can lead to some interesting new insights, such as described below.

How can a very uniform electric field be created? Consider a system of two metal plates with opposite charges on them, as shown in [link]. The properties of conductors in electrostatic equilibrium indicate that the electric field between the plates will be uniform in strength and direction. Except near the edges, the excess charges distribute themselves uniformly, producing field lines that are uniformly spaced (hence uniform in strength) and perpendicular to the surfaces (hence uniform in direction, since the plates are flat). The edge effects are less important when the plates are close together.



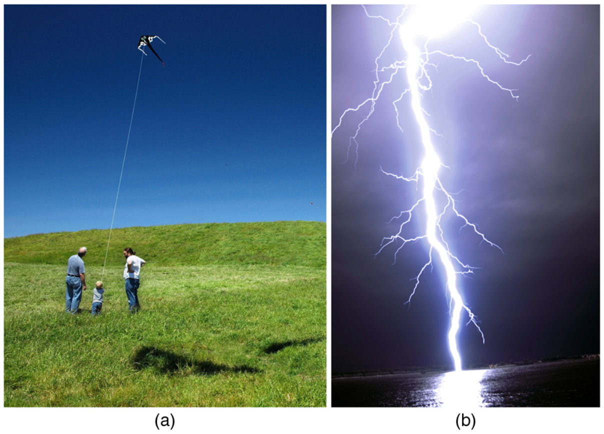
Two metal plates with equal, but opposite, excess charges. The field between them is uniform in strength and direction except near the edges. One use of such a field is to produce uniform acceleration of charges between the plates, such as in the electron gun of a TV tube.

Earth’s Electric Field

A near uniform electric field of approximately 150 N/C, directed downward, surrounds Earth, with the magnitude increasing slightly as we get closer to the surface. What causes the electric field? At around 100 km above the surface of Earth we have a layer of charged particles, called the **ionosphere**. The ionosphere is responsible for a range of phenomena including the electric field surrounding Earth. In fair weather the ionosphere is positive and the Earth largely negative, maintaining the electric field ([link](a)).

In storm conditions clouds form and localized electric fields can be larger and reversed in direction ([link](b)). The exact charge distributions depend on the local conditions, and variations of [link](b) are possible.

If the electric field is sufficiently large, the insulating properties of the surrounding material break down and it becomes conducting. For air this occurs at around 3×10^6 N/C. Air ionizes ions and electrons recombine, and we get discharge in the form of lightning sparks and corona discharge.



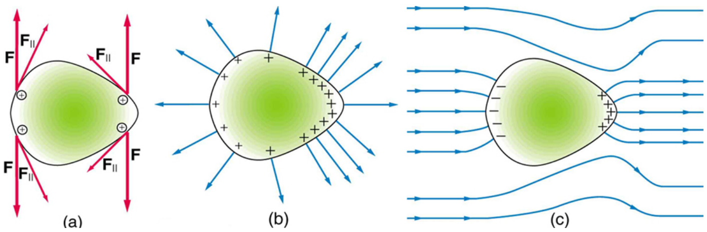
Earth’s electric field. (a) Fair weather field. Earth and the ionosphere (a layer of charged particles) are both conductors. They produce a uniform electric field of about 150 N/C. (credit: D. H. Parks) (b) Storm fields. In the presence of storm clouds, the local electric fields can be larger. At very high fields, the insulating properties of the air break down and lightning can occur. (credit: Jan-Joost Verhoef)

Electric Fields on Uneven Surfaces

So far we have considered excess charges on a smooth, symmetrical conductor surface. What happens if a conductor has sharp corners or is pointed? Excess charges on a nonuniform conductor become concentrated at the sharpest points. Additionally, excess charge may move on or off the conductor at the sharpest points.

To see how and why this happens, consider the charged conductor in [link]. The electrostatic repulsion of like charges is most effective in moving them apart on the flattest surface, and so they become least concentrated there. This is because the forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surfaces are different. The component parallel to the surface is greatest on the flattest surface and, hence, more effective in moving the charge.

The same effect is produced on a conductor by an externally applied electric field, as seen in [link] (c). Since the field lines must be perpendicular to the surface, more of them are concentrated on the most curved parts.



Excess charge on a nonuniform conductor becomes most concentrated at the location of greatest curvature. (a) The forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surface are different. It is $\mathbf{F}_{||}$ that moves the charges apart once they have reached the surface. (b) $\mathbf{F}_{||}$ is smallest at the more pointed end, the charges are left closer together, producing the electric field shown. (c) An uncharged conductor in an originally uniform electric field is polarized, with the most concentrated charge at its most pointed end.

Applications of Conductors

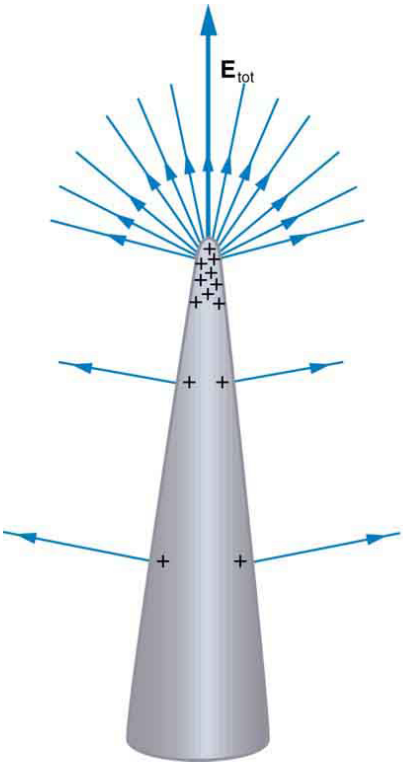
On a very sharply curved surface, such as shown in [link], the charges are so concentrated at the point that the resulting electric field can be great enough to remove them from the surface. This can be useful.

Lightning rods work best when they are most pointed. The large charges created in storm clouds induce an opposite charge on a building that can result in a lightning bolt hitting the building. The induced charge is bled away continually by a lightning rod, preventing the more dramatic lightning strike.

Of course, we sometimes wish to prevent the transfer of charge rather than to facilitate it. In that case, the conductor should be very smooth and have as large a radius of curvature as possible. (See [link].) Smooth surfaces are used on high-voltage transmission lines, for example, to avoid leakage of charge into the air.

Another device that makes use of some of these principles is a **Faraday cage**. This is a metal shield that encloses a volume. All electrical charges will reside on the outside surface of this shield, and there will be no electrical field inside. A Faraday cage is used to prohibit stray electrical fields in the environment from interfering with sensitive measurements, such as the electrical signals inside a nerve cell.

During electrical storms if you are driving a car, it is best to stay inside the car as its metal body acts as a Faraday cage with zero electrical field inside. If in the vicinity of a lightning strike, its effect is felt on the outside of the car and the inside is unaffected, provided you remain totally inside. This is also true if an active (“hot”) electrical wire was broken (in a storm or an accident) and fell on your car.



A very pointed conductor has a large charge concentration at the point. The electric field is very strong at the point and can exert a force large enough to transfer charge on or off the conductor.

Lightning rods are used to prevent the buildup of large excess charges on structures and, thus, are pointed.

(a) A lightning rod is pointed to facilitate the transfer of charge. (credit: Romaine, Wikimedia Commons) (b) This Van de Graaff generator has a smooth surface with a large radius of curvature to prevent the transfer of charge and allow a large voltage to be generated. The mutual repulsion of like charges is evident in the person’s hair while touching the metal sphere. (credit: Jon ‘ShakataGaNai’ Davis/Wikimedia Commons).

Section Summary

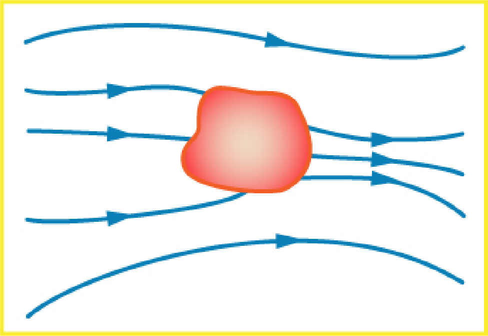
- A conductor allows free charges to move about within it.
- The electrical forces around a conductor will cause free charges to move around inside the conductor until static equilibrium is reached.
- Any excess charge will collect along the surface of a conductor.
- Conductors with sharp corners or points will collect more charge at those points.
- A lightning rod is a conductor with sharply pointed ends that collect excess charge on the building caused by an electrical storm and allow it to dissipate back into the air.
- Electrical storms result when the electrical field of Earth’s surface in certain locations becomes more strongly charged, due to changes in the insulating effect of the air.
- A Faraday cage acts like a shield around an object, preventing electric charge from penetrating inside.

Conceptual Questions

Exercise:

Problem:

Is the object in [link] a conductor or an insulator? Justify your answer.



Exercise:

Problem:

If the electric field lines in the figure above were perpendicular to the object, would it necessarily be a conductor? Explain.

Exercise:

Problem:

The discussion of the electric field between two parallel conducting plates, in this module states that edge effects are less important if the plates are close together. What does close mean? That is, is the actual plate separation crucial, or is the ratio of plate separation to plate area crucial?

Exercise:

Problem:

Would the self-created electric field at the end of a pointed conductor, such as a lightning rod, remove positive or negative charge from the conductor? Would the same sign charge be removed from a neutral pointed conductor by the application of a similar externally created electric field? (The answers to both questions have implications for charge transfer utilizing points.)

Exercise:

Problem:

Why is a golfer with a metal club over her shoulder vulnerable to lightning in an open fairway? Would she be any safer under a tree?

Exercise:

Problem:

Can the belt of a Van de Graaff accelerator be a conductor? Explain.

Exercise:

Problem:

Are you relatively safe from lightning inside an automobile? Give two reasons.

Exercise:

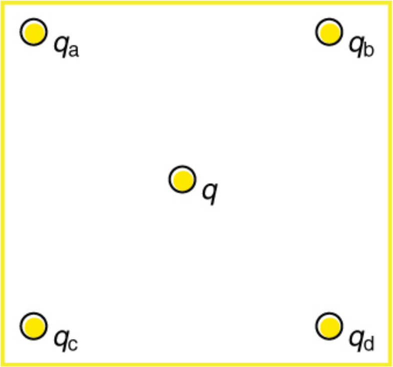
Problem:

Discuss pros and cons of a lightning rod being grounded versus simply being attached to a building.

Exercise:

Problem:

Using the symmetry of the arrangement, show that the net Coulomb force on the charge *q* at the center of the square below ([link]) is zero if the charges on the four corners are exactly equal.



Four point charges q_a, q_b, q_c, and q_d lie on the corners of a square and *q* is located at its center.

Exercise:

Problem:

(a) Using the symmetry of the arrangement, show that the electric field at the center of the square in [link] is zero if the charges on the four corners are exactly equal. (b) Show that this is also true for any combination of charges in which $q_a = q_d$ and $q_b = q_c$

Exercise:

Problem:

(a) What is the direction of the total Coulomb force on q in [link] if q is negative, $q_a = q_c$ and both are negative, and $q_b = q_c$ and both are positive? (b) What is the direction of the electric field at the center of the square in this situation?

Exercise:

Problem:

Considering [link], suppose that $q_a = q_d$ and $q_b = q_c$. First show that q is in static equilibrium. (You may neglect the gravitational force.) Then discuss whether the equilibrium is stable or unstable, noting that this may depend on the signs of the charges and the direction of displacement of q from the center of the square.

Exercise:

Problem:

If $q_a = 0$ in [link], under what conditions will there be no net Coulomb force on q?

Exercise:

Problem:

In regions of low humidity, one develops a special “grip” when opening car doors, or touching metal door knobs. This involves placing as much of the hand on the device as possible, not just the ends of one’s fingers. Discuss the induced charge and explain why this is done.

Exercise:

Problem:

Tollbooth stations on roadways and bridges usually have a piece of wire stuck in the pavement before them that will touch a car as it approaches. Why is this done?

Exercise:

Problem:

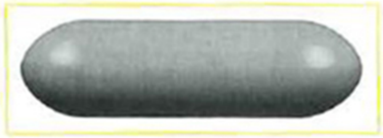
Suppose a woman carries an excess charge. To maintain her charged status can she be standing on ground wearing just any pair of shoes? How would you discharge her? What are the consequences if she simply walks away?

Problems & Exercises

Exercise:

Problem:

Sketch the electric field lines in the vicinity of the conductor in [link] given the field was originally uniform and parallel to the object’s long axis. Is the resulting field small near the long side of the object?



Exercise:

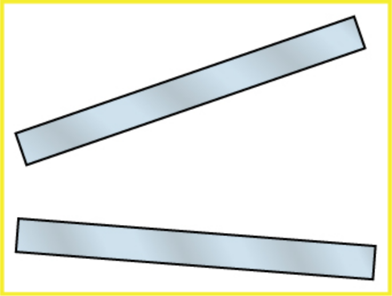
Problem:

Sketch the electric field lines in the vicinity of the conductor in [link] given the field was originally uniform and parallel to the object’s long axis. Is the resulting field small near the long side of the object?

Exercise:

Problem:

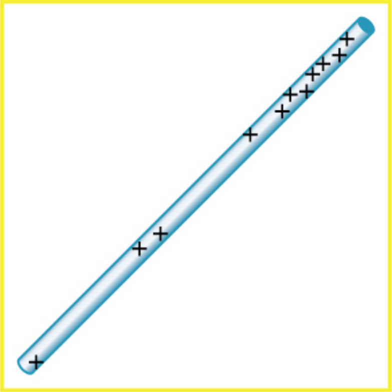
Sketch the electric field between the two conducting plates shown in [link], given the top plate is positive and an equal amount of negative charge is on the bottom plate. Be certain to indicate the distribution of charge on the plates.



Exercise:

Problem:

Sketch the electric field lines in the vicinity of the charged insulator in [link] noting its nonuniform charge distribution.

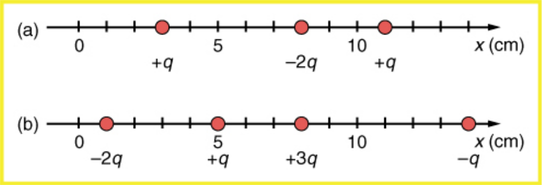


A charged insulating rod such as might be used in a classroom demonstration.

Exercise:

Problem:

What is the force on the charge located at $x = 8.00\text{ cm}$ in [link](a) given that $q = 1.00\ \mu\text{C}$?



(a) Point charges located at 3.00, 8.00, and 11.0 cm along the *x*-axis. (b) Point charges located at 1.00, 5.00, 8.00, and 14.0 cm along the *x*-axis.

Exercise:

Problem:

(a) Find the total electric field at $x = 1.00\text{ cm}$ in [link](b) given that $q = 5.00\text{ nC}$. (b) Find the total electric field at $x = 11.00\text{ cm}$ in [link](b). (c) If the charges are allowed to move and eventually be brought to rest by friction, what will the final charge configuration be? (That is, will there be a single charge, double charge, etc., and what will its value(s) be?)

Solution:

(a) $E_{x=1.00\text{ cm}} = -\infty$

(b) $2.12\times 10^{5}\ \text{N/C}$

(c) one charge of $+q$

Exercise:

Problem:

(a) Find the electric field at $x=5.00\ \text{cm}$ in [link](a), given that $q=1.00\ \text{µC}$. (b) At what position between 3.00 and 8.00 cm is the total electric field the same as that for –2q alone? (c) Can the electric field be zero anywhere between 0.00 and 8.00 cm? (d) At very large positive or negative values of *x*, the electric field approaches zero in both (a) and (b). In which does it most rapidly approach zero and why? (e) At what position to the right of 11.0 cm is the total electric field zero, other than at infinity? (Hint: A graphing calculator can yield considerable insight in this problem.)

Exercise:

Problem:

(a) Find the total Coulomb force on a charge of 2.00 nC located at $x=4.00\ \text{cm}$ in [link] (b), given that $q=1.00\ \text{µC}$. (b) Find the *x*-position at which the electric field is zero in [link] (b).

Solution:

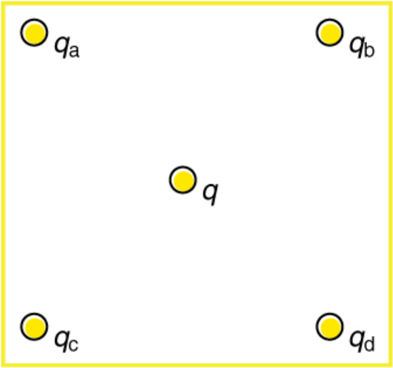
(a) 0.252 N to the left

(b) $x=6.07\ \text{cm}$

Exercise:

Problem:

Using the symmetry of the arrangement, determine the direction of the force on q in the figure below, given that $q_a=q_b$=+7.50 µC and $q_c=q_d=-7.50\ \text{µC}$. (b) Calculate the magnitude of the force on the charge q, given that the square is 10.0 cm on a side and $q=2.00\ \text{µC}$.



Exercise:

Problem:

(a) Using the symmetry of the arrangement, determine the direction of the electric field at the center of the square in [link], given that $q_a = q_b = -1.00\ \mu\text{C}$ and $q_c = q_d$=$+1.00\ \mu\text{C}$. (b) Calculate the magnitude of the electric field at the location of q, given that the square is 5.00 cm on a side.

Solution:

(a)The electric field at the center of the square will be straight up, since q_a and q_b are positive and q_c and q_d are negative and all have the same magnitude.

(b) $2.04 \times 10^7\ \text{N/C (upward)}$

Exercise:

Problem:

Find the electric field at the location of q_a in [link] given that $q_b = q_c = q_d$=$+2.00\ \text{nC}$, $q = -1.00\ \text{nC}$, and the square is 20.0 cm on a side.

Exercise:

Problem:

Find the total Coulomb force on the charge q in [link], given that $q = 1.00\ \mu\text{C}$, $q_a = 2.00\ \mu\text{C}$, $q_b = -3.00\ \mu\text{C}$, $q_c = -4.00\ \mu\text{C}$, and q_d=$+1.00\ \mu\text{C}$. The square is 50.0 cm on a side.

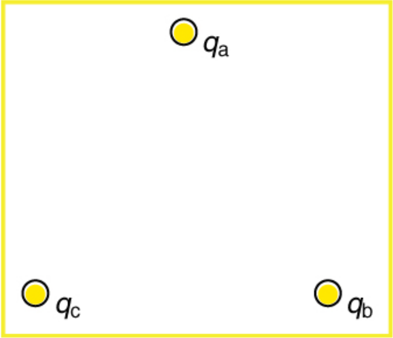
Solution:

$0.102\ \text{N}$, in the $-y$ direction

Exercise:

Problem:

(a) Find the electric field at the location of q_a in [link], given that $q_\text{b} = +10.00\ \mu\text{C}$ and $q_\text{c} = –5.00\ \mu\text{C}$. (b) What is the force on q_a, given that $q_\text{a} = +1.50\ \text{nC}$?



Point charges located at the corners of an equilateral triangle 25.0 cm on a side.

Exercise:

Problem:

(a) Find the electric field at the center of the triangular configuration of charges in [link], given that q_a=$+2.50\text{ nC}$, $q_b = -8.00\text{ nC}$, and q_c=$+1.50\text{ nC}$. (b) Is there any combination of charges, other than $q_a = q_b = q_c$, that will produce a zero strength electric field at the center of the triangular configuration?

Solution:

(a) $E = 4.36 \times 10^3\text{ N/C},\,35.0$º, below the horizontal.

(b) No

Glossary

conductor
: an object with properties that allow charges to move about freely within it

free charge
: an electrical charge (either positive or negative) which can move about separately from its base molecule

electrostatic equilibrium
: an electrostatically balanced state in which all free electrical charges have stopped moving about

polarized
: a state in which the positive and negative charges within an object have collected in separate locations

ionosphere
: a layer of charged particles located around 100 km above the surface of Earth, which is responsible for a range of phenomena including the electric field surrounding Earth

Faraday cage

a metal shield which prevents electric charge from penetrating its surface

Applications of Electrostatics

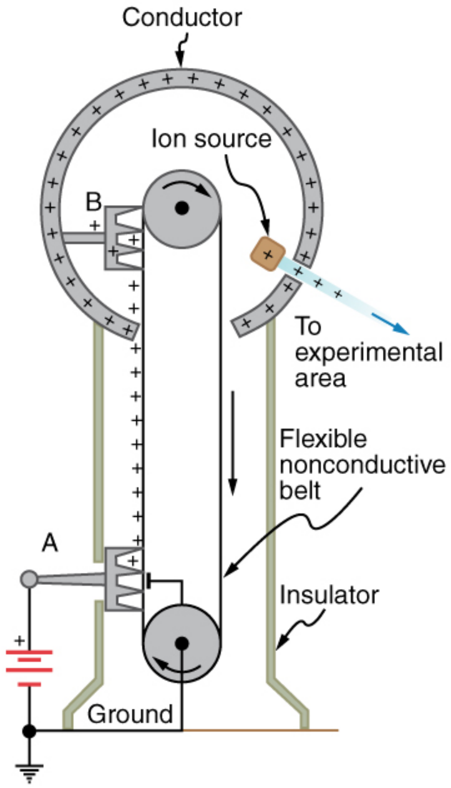
- Name several real-world applications of the study of electrostatics.

The study of **electrostatics** has proven useful in many areas. This module covers just a few of the many applications of electrostatics.

The Van de Graaff Generator

Van de Graaff generators (or Van de Graaffs) are not only spectacular devices used to demonstrate high voltage due to static electricity—they are also used for serious research. The first was built by Robert Van de Graaff in 1931 (based on original suggestions by Lord Kelvin) for use in nuclear physics research. [link] shows a schematic of a large research version. Van de Graaffs utilize both smooth and pointed surfaces, and conductors and insulators to generate large static charges and, hence, large voltages.

A very large excess charge can be deposited on the sphere, because it moves quickly to the outer surface. Practical limits arise because the large electric fields polarize and eventually ionize surrounding materials, creating free charges that neutralize excess charge or allow it to escape. Nevertheless, voltages of 15 million volts are well within practical limits.



Schematic of Van de Graaff generator. A battery (A) supplies excess positive charge to a pointed conductor, the points of which spray the charge onto a moving insulating belt near the bottom. The pointed conductor (B) on top in the large sphere picks up the charge. (The induced electric field at the points is so large that it removes the charge from the belt.) This can be done because the charge does not remain inside the conducting sphere but moves to its outside surface. An ion source inside the sphere produces positive ions, which are accelerated away from the positive sphere to high velocities.

Note:
Take-Home Experiment: Electrostatics and Humidity
Rub a comb through your hair and use it to lift pieces of paper. It may help to tear the pieces of paper rather than cut them neatly. Repeat the exercise in your bathroom after you have had a long shower and the air in the bathroom is moist. Is it easier to get electrostatic effects in dry or moist air? Why would torn paper be more attractive to the comb than cut paper? Explain your observations.

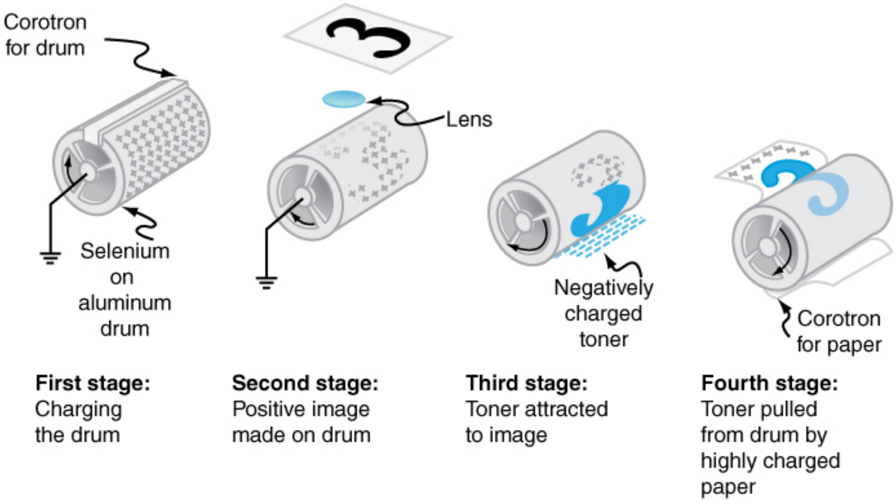
Xerography

Most copy machines use an electrostatic process called **xerography**—a word coined from the Greek words *xeros* for dry and *graphos* for writing. The heart of the process is shown in simplified form in [link].

A selenium-coated aluminum drum is sprayed with positive charge from points on a device called a corotron. Selenium is a substance with an interesting property—it is a **photoconductor**. That is, selenium is an insulator when in the dark and a conductor when exposed to light.

In the first stage of the xerography process, the conducting aluminum drum is **grounded** so that a negative charge is induced under the thin layer of uniformly positively charged selenium. In the second stage, the surface of the drum is exposed to the image of whatever is to be copied. Where the image is light, the selenium becomes conducting, and the positive charge is neutralized. In dark areas, the positive charge remains, and so the image has been transferred to the drum.

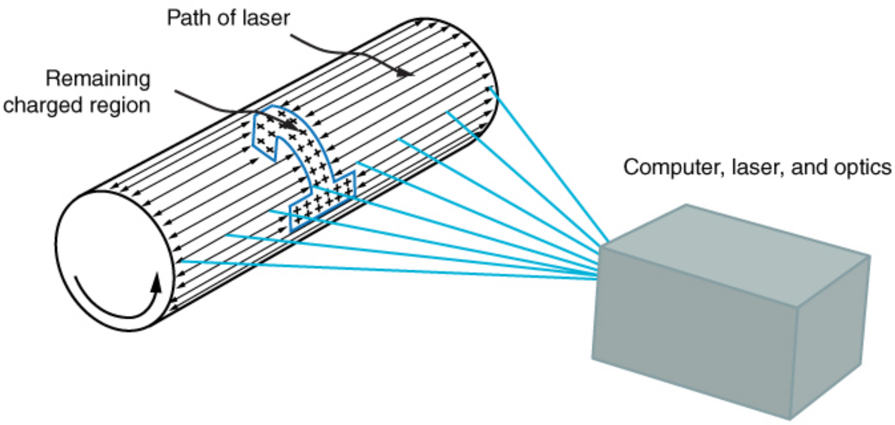
The third stage takes a dry black powder, called toner, and sprays it with a negative charge so that it will be attracted to the positive regions of the drum. Next, a blank piece of paper is given a greater positive charge than on the drum so that it will pull the toner from the drum. Finally, the paper and electrostatically held toner are passed through heated pressure rollers, which melt and permanently adhere the toner within the fibers of the paper.



Xerography is a dry copying process based on electrostatics. The major steps in the process are the charging of the photoconducting drum, transfer of an image creating a positive charge duplicate, attraction of toner to the charged parts of the drum, and transfer of toner to the paper. Not shown are heat treatment of the paper and cleansing of the drum for the next copy.

Laser Printers

Laser printers use the xerographic process to make high-quality images on paper, employing a laser to produce an image on the photoconducting drum as shown in [link]. In its most common application, the laser printer receives output from a computer, and it can achieve high-quality output because of the precision with which laser light can be controlled. Many laser printers do significant information processing, such as making sophisticated letters or fonts, and may contain a computer more powerful than the one giving them the raw data to be printed.

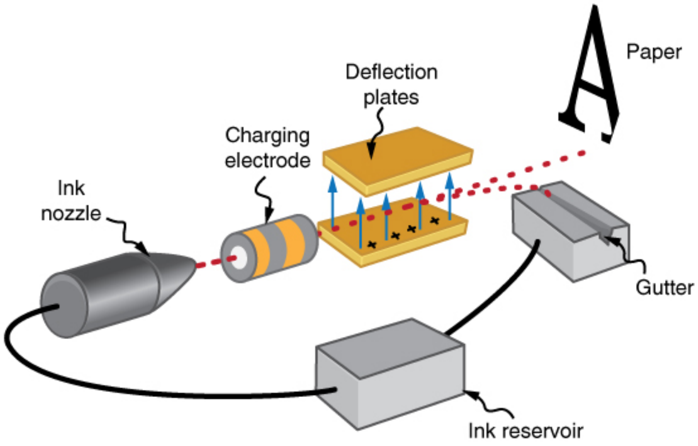


In a laser printer, a laser beam is scanned across a photoconducting drum, leaving a positive charge image. The other steps for charging the drum and transferring the image to paper are the same as in xerography. Laser light can be very precisely controlled, enabling laser printers to produce high-quality images.

Ink Jet Printers and Electrostatic Painting

The **ink jet printer**, commonly used to print computer-generated text and graphics, also employs electrostatics. A nozzle makes a fine spray of tiny ink droplets, which are then given an electrostatic charge. (See [link].)

Once charged, the droplets can be directed, using pairs of charged plates, with great precision to form letters and images on paper. Ink jet printers can produce color images by using a black jet and three other jets with primary colors, usually cyan, magenta, and yellow, much as a color television produces color. (This is more difficult with xerography, requiring multiple drums and toners.)



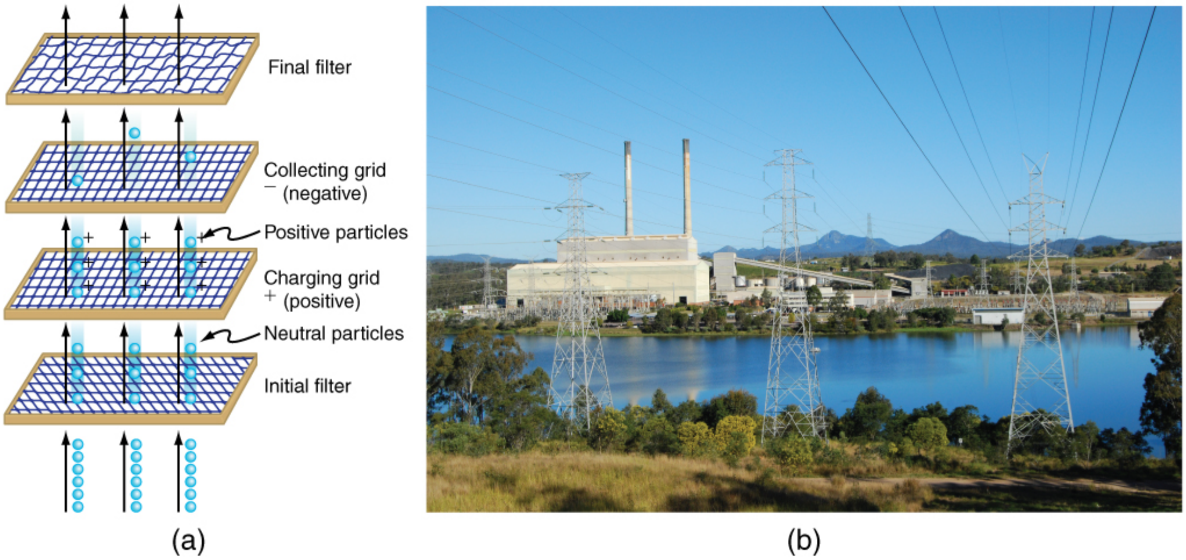
The nozzle of an ink-jet printer produces small ink droplets, which are sprayed with electrostatic charge. Various computer-driven devices are then used to direct the droplets to the correct positions on a page.

Electrostatic painting employs electrostatic charge to spray paint onto odd-shaped surfaces. Mutual repulsion of like charges causes the paint to fly away from its source. Surface tension forms drops, which are then attracted by unlike charges to the surface to be painted. Electrostatic painting can reach those hard-to-get at places, applying an even coat in a controlled manner. If the object is a conductor, the electric field is perpendicular to the surface, tending to bring the drops in perpendicularly. Corners and points on conductors will receive extra paint. Felt can similarly be applied.

Smoke Precipitators and Electrostatic Air Cleaning

Another important application of electrostatics is found in air cleaners, both large and small. The electrostatic part of the process places excess (usually positive) charge on smoke, dust, pollen, and other particles in the air and then passes the air through an oppositely charged grid that attracts and retains the charged particles. (See [link].)

Large **electrostatic precipitators** are used industrially to remove over 99% of the particles from stack gas emissions associated with the burning of coal and oil. Home precipitators, often in conjunction with the home heating and air conditioning system, are very effective in removing polluting particles, irritants, and allergens.



(a) Schematic of an electrostatic precipitator. Air is passed through grids of opposite charge. The first grid charges airborne particles, while the second attracts and collects them. (b) The dramatic effect of electrostatic precipitators is seen by the absence of smoke from this power plant. (credit: Cmdalgleish, Wikimedia Commons)

Note:
Problem-Solving Strategies for Electrostatics

1. Examine the situation to determine if static electricity is involved. This may concern separated stationary charges, the forces among them, and the electric fields they create.
2. Identify the system of interest. This includes noting the number, locations, and types of charges involved.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful. Determine whether the Coulomb force is to be considered directly—if so, it may be useful to draw a free-body diagram, using electric field lines.
4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns). It is important to distinguish the Coulomb force F from the electric field E, for example.
5. Solve the appropriate equation for the quantity to be determined (the unknown) or draw the field lines as requested.
6. Examine the answer to see if it is reasonable: Does it make sense? Are units correct and the numbers involved reasonable?

Integrated Concepts

The Integrated Concepts exercises for this module involve concepts such as electric charges, electric fields, and several other topics. Physics is most interesting when applied to general situations involving more than a narrow set of physical principles. The electric field exerts force on charges, for example, and hence the relevance of Dynamics: Force and Newton’s Laws of Motion. The following topics are involved in some or all of the problems labeled “Integrated Concepts”:

- Kinematics
- Two-Dimensional Kinematics
- Dynamics: Force and Newton’s Laws of Motion
- Uniform Circular Motion and Gravitation
- Statics and Torque
- Fluid Statics

The following worked example illustrates how this strategy is applied to an Integrated Concept problem:

Example:

Acceleration of a Charged Drop of Gasoline

If steps are not taken to ground a gasoline pump, static electricity can be placed on gasoline when filling your car’s tank. Suppose a tiny drop of gasoline has a mass of 4.00×10^{-15} kg and is given a positive charge of 3.20×10^{-19} C. (a) Find the weight of the drop. (b) Calculate the electric force on the drop if there is an upward electric field of strength 3.00×10^{5} N/C due to other static electricity in the vicinity. (c) Calculate the drop’s acceleration.

Strategy

To solve an integrated concept problem, we must first identify the physical principles involved and identify the chapters in which they are found. Part (a) of this example asks for weight. This is a topic of dynamics and is defined in Dynamics: Force and Newton’s Laws of Motion. Part (b) deals with electric force on a charge, a topic of Electric Charge and Electric Field. Part (c) asks for acceleration, knowing forces and mass. These are part of Newton’s laws, also found in Dynamics: Force and Newton’s Laws of Motion.

The following solutions to each part of the example illustrate how the specific problem-solving strategies are applied. These involve identifying knowns and unknowns, checking to see if the answer is reasonable, and so on.

Solution for (a)

Weight is mass times the acceleration due to gravity, as first expressed in

Equation:

$$w = \text{mg}.$$

Entering the given mass and the average acceleration due to gravity yields

Equation:

$$w = (4.00 \times 10^{-15} \text{ kg})(9.80 \text{ m/s}^2) = 3.92 \times 10^{-14} \text{ N}.$$

Discussion for (a)

This is a small weight, consistent with the small mass of the drop.

Solution for (b)

The force an electric field exerts on a charge is given by rearranging the following equation:

Equation:

$$F = \text{qE}.$$

Here we are given the charge (3.20×10^{-19} C is twice the fundamental unit of charge) and the electric field strength, and so the electric force is found to be

Equation:

$$F = (3.20 \times 10^{-19} \text{ C})(3.00 \times 10^5 \text{ N/C}) = 9.60 \times 10^{-14} \text{ N}.$$

Discussion for (b)

While this is a small force, it is greater than the weight of the drop.

Solution for (c)

The acceleration can be found using Newton’s second law, provided we can identify all of the external forces acting on the drop. We assume only the drop’s weight and the electric force are significant. Since the drop has a positive charge and the electric field is given to be upward, the electric force is upward. We thus have a one-dimensional (vertical direction) problem, and we can state Newton’s second law as

Equation:

$$a = \frac{F_{\text{net}}}{m}.$$

where $F_{\text{net}} = F - w$. Entering this and the known values into the expression for Newton's second law yields

Equation:

$$\begin{aligned} a &= \frac{F-w}{m} \\ &= \frac{9.60\times10^{-14}\text{ N}-3.92\times10^{-14}\text{ N}}{4.00\times10^{-15}\text{ kg}} \\ &= 14.2\text{ m/s}^2. \end{aligned}$$

Discussion for (c)

This is an upward acceleration great enough to carry the drop to places where you might not wish to have gasoline.

This worked example illustrates how to apply problem-solving strategies to situations that include topics in different chapters. The first step is to identify the physical principles involved in the problem. The second step is to solve for the unknown using familiar problem-solving strategies. These are found throughout the text, and many worked examples show how to use them for single topics. In this integrated concepts example, you can see how to apply them across several topics. You will find these techniques useful in applications of physics outside a physics course, such as in your profession, in other science disciplines, and in everyday life. The following problems will build your skills in the broad application of physical principles.

Note:

Unreasonable Results

The Unreasonable Results exercises for this module have results that are unreasonable because some premise is unreasonable or because certain of the premises are inconsistent with one another. Physical principles applied correctly then produce unreasonable results. The purpose of these problems is to give practice in assessing whether nature is being accurately described, and if it is not to trace the source of difficulty.

Note:

Problem-Solving Strategy

To determine if an answer is reasonable, and to determine the cause if it is not, do the following.

1. Solve the problem using strategies as outlined above. Use the format followed in the worked examples in the text to solve the problem as usual.
2. Check to see if the answer is reasonable. Is it too large or too small, or does it have the wrong sign, improper units, and so on?
3. If the answer is unreasonable, look for what specifically could cause the identified difficulty. Usually, the manner in which the answer is unreasonable is an indication of the difficulty. For example, an extremely large Coulomb force could be due to the assumption of an excessively large separated charge.

Section Summary

- Electrostatics is the study of electric fields in static equilibrium.
- In addition to research using equipment such as a Van de Graaff generator, many practical applications of electrostatics exist, including photocopiers, laser printers, ink-jet printers and electrostatic air filters.

Problems & Exercises

Exercise:

Problem:

(a) What is the electric field 5.00 m from the center of the terminal of a Van de Graaff with a 3.00 mC charge, noting that the field is equivalent to that of a point charge at the center of the terminal? (b) At this distance, what force does the field exert on a $2.00\ \mu\text{C}$ charge on the Van de Graaff's belt?

Exercise:

Problem:

(a) What is the direction and magnitude of an electric field that supports the weight of a free electron near the surface of Earth? (b) Discuss what the small value for this field implies regarding the relative strength of the gravitational and electrostatic forces.

Solution:

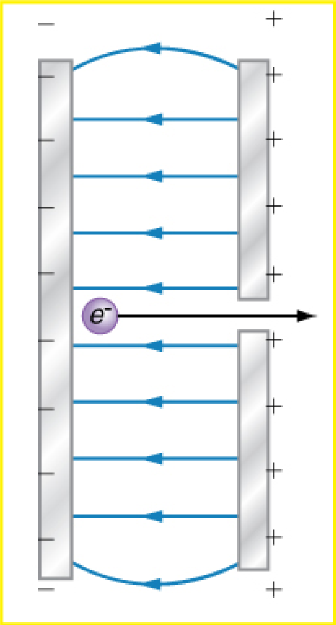
(a) $5.58\times10^{-11}\ \text{N/C}$

(b)the coulomb force is extraordinarily stronger than gravity

Exercise:

Problem:

A simple and common technique for accelerating electrons is shown in [link], where there is a uniform electric field between two plates. Electrons are released, usually from a hot filament, near the negative plate, and there is a small hole in the positive plate that allows the electrons to continue moving. (a) Calculate the acceleration of the electron if the field strength is $2.50\times10^{4}\ \text{N/C}$. (b) Explain why the electron will not be pulled back to the positive plate once it moves through the hole.



Parallel conducting plates with opposite charges on them create a relatively uniform electric field used to accelerate electrons to the right. Those that go through the hole can be used to make a TV or computer screen glow or to produce X-rays.

Exercise:

Problem:

Earth has a net charge that produces an electric field of approximately 150 N/C downward at its surface. (a) What is the magnitude and sign of the excess charge, noting the electric field of a conducting sphere is equivalent to a point charge at its center? (b) What acceleration will the field produce on a free electron near Earth's surface? (c) What mass object with a single extra electron will have its weight supported by this field?

Solution:

(a) $-6.76 \times 10^{5}\ \text{C}$

(b) $2.63 \times 10^{13}\ \text{m/s}^{2}\ \text{(upward)}$

(c) $2.45 \times 10^{-18}\ \text{kg}$

Exercise:

Problem:

Point charges of $25.0\ \mu\text{C}$ and $45.0\ \mu\text{C}$ are placed 0.500 m apart. (a) At what point along the line between them is the electric field zero? (b) What is the electric field halfway between them?

Exercise:

Problem:

What can you say about two charges q_1 and q_2, if the electric field one-fourth of the way from q_1 to q_2 is zero?

Solution:

The charge q_2 is 9 times greater than q_1.

Exercise:

Problem: **Integrated Concepts**

Calculate the angular velocity ω of an electron orbiting a proton in the hydrogen atom, given the radius of the orbit is 0.530×10^{-10} m. You may assume that the proton is stationary and the centripetal force is supplied by Coulomb attraction.

Exercise:

Problem: Integrated Concepts

An electron has an initial velocity of $5.00 \times 10^{6}\ \text{m/s}$ in a uniform $2.00 \times 10^{5}\ \text{N/C}$ strength electric field. The field accelerates the electron in the direction opposite to its initial velocity. (a) What is the direction of the electric field? (b) How far does the electron travel before coming to rest? (c) How long does it take the electron to come to rest? (d) What is the electron's velocity when it returns to its starting point?

Exercise:

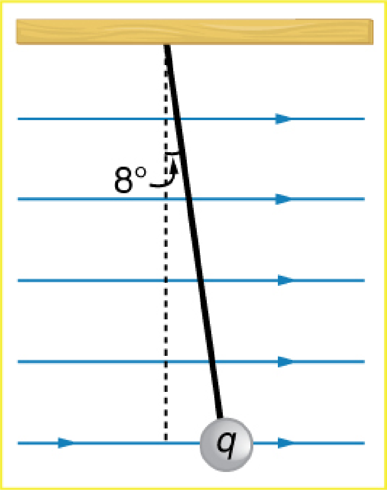
Problem: Integrated Concepts

The practical limit to an electric field in air is about $3.00 \times 10^{6}\ \text{N/C}$. Above this strength, sparking takes place because air begins to ionize and charges flow, reducing the field. (a) Calculate the distance a free proton must travel in this field to reach 3.00% of the speed of light, starting from rest. (b) Is this practical in air, or must it occur in a vacuum?

Exercise:

Problem: Integrated Concepts

A 5.00 g charged insulating ball hangs on a 30.0 cm long string in a uniform horizontal electric field as shown in [link]. Given the charge on the ball is $1.00\ \mu\text{C}$, find the strength of the field.

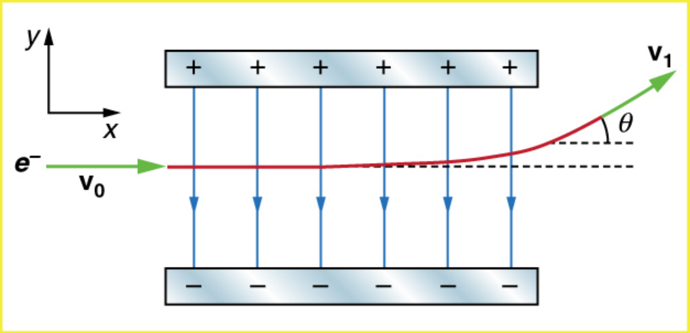


A horizontal electric field causes the charged ball to hang at an angle of 8.00º.

Exercise:

Problem: Integrated Concepts

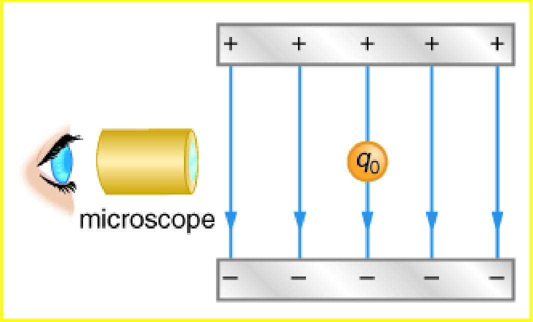
[link] shows an electron passing between two charged metal plates that create an 100 N/C vertical electric field perpendicular to the electron’s original horizontal velocity. (These can be used to change the electron’s direction, such as in an oscilloscope.) The initial speed of the electron is $3.00 \times 10^6 \text{ m/s}$, and the horizontal distance it travels in the uniform field is 4.00 cm. (a) What is its vertical deflection? (b) What is the vertical component of its final velocity? (c) At what angle does it exit? Neglect any edge effects.



Exercise:

Problem: Integrated Concepts

The classic Millikan oil drop experiment was the first to obtain an accurate measurement of the charge on an electron. In it, oil drops were suspended against the gravitational force by a vertical electric field. (See [link].) Given the oil drop to be $1.00\ \mu\text{m}$ in radius and have a density of $920\ \text{kg/m}^3$: (a) Find the weight of the drop. (b) If the drop has a single excess electron, find the electric field strength needed to balance its weight.

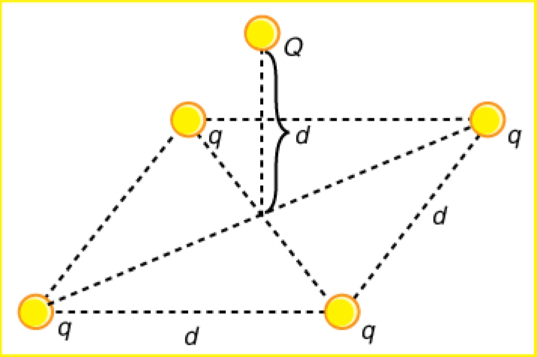


In the Millikan oil drop experiment, small drops can be suspended in an electric field by the force exerted on a single excess electron. Classically, this experiment was used to determine the electron charge q_e by measuring the electric field and mass of the drop.

Exercise:

Problem: Integrated Concepts

(a) In [link], four equal charges q lie on the corners of a square. A fifth charge Q is on a mass m directly above the center of the square, at a height equal to the length d of one side of the square. Determine the magnitude of q in terms of Q, m, and d, if the Coulomb force is to equal the weight of m. (b) Is this equilibrium stable or unstable? Discuss.



Four equal charges on the corners of a horizontal square support the weight of a fifth charge located directly above the center of the square.

Exercise:

Problem: Unreasonable Results

(a) Calculate the electric field strength near a 10.0 cm diameter conducting sphere that has 1.00 C of excess charge on it. (b) What is unreasonable about this result? (c) Which assumptions are responsible?

Exercise:

Problem: Unreasonable Results

(a) Two 0.500 g raindrops in a thunderhead are 1.00 cm apart when they each acquire 1.00 mC charges. Find their acceleration. (b) What is unreasonable about this result? (c) Which premise or assumption is responsible?

Exercise:

Problem: Unreasonable Results

A wrecking yard inventor wants to pick up cars by charging a 0.400 m diameter ball and inducing an equal and opposite charge on the car. If a car has a 1000 kg mass and the ball is to be able to lift it from a distance of 1.00 m: (a) What minimum charge must be used? (b) What is the electric field near the surface of the ball? (c) Why are these results unreasonable? (d) Which premise or assumption is responsible?

Exercise:

Problem: Construct Your Own Problem

Consider two insulating balls with evenly distributed equal and opposite charges on their surfaces, held with a certain distance between the centers of the balls. Construct a problem in which you calculate the electric field (magnitude and direction) due to the balls at various points along a line running through the centers of the balls and extending to infinity on either side. Choose interesting points and comment on the meaning of the field at those points. For example, at what points might the field be just that due to one ball and where does the field become negligibly small? Among the things to be considered are the magnitudes of the charges and the distance between the centers of the balls. Your instructor may wish for you to consider the electric field off axis or for a more complex array of charges, such as those in a water molecule.

Exercise:

Problem: Construct Your Own Problem

Consider identical spherical conducting space ships in deep space where gravitational fields from other bodies are negligible compared to the gravitational attraction between the ships. Construct a problem in which you place identical excess charges on the space ships to exactly counter their gravitational attraction. Calculate the amount of excess charge needed. Examine whether that charge depends on the distance between the centers of the ships, the masses of the ships, or any other factors. Discuss whether this would be an easy, difficult, or even impossible thing to do in practice.

Glossary

Van de Graaff generator
: a machine that produces a large amount of excess charge, used for experiments with high voltage

electrostatics
: the study of electric forces that are static or slow-moving

photoconductor
: a substance that is an insulator until it is exposed to light, when it becomes a conductor

xerography
: a dry copying process based on electrostatics

grounded
: connected to the ground with a conductor, so that charge flows freely to and from the Earth to the grounded object

laser printer

uses a laser to create a photoconductive image on a drum, which attracts dry ink particles that are then rolled onto a sheet of paper to print a high-quality copy of the image

ink-jet printer

small ink droplets sprayed with an electric charge are controlled by electrostatic plates to create images on paper

electrostatic precipitators

filters that apply charges to particles in the air, then attract those charges to a filter, removing them from the airstream

Introduction to Electromagnetic Waves
class="introduction"

Human eyes detect these orange “sea goldie” fish swimming over a coral reef in the blue waters of the Gulf of Eilat (Red Sea) using visible light. (credit: Daviddarom, Wikimedia Commons)

The beauty of a coral reef, the warm radiance of sunshine, the sting of sunburn, the X-ray revealing a broken bone, even microwave popcorn—all are brought to us by **electromagnetic waves**. The list of the various types of electromagnetic waves, ranging from radio transmission waves to nuclear gamma-ray (γ-ray) emissions, is interesting in itself.

Even more intriguing is that all of these widely varied phenomena are different manifestations of the same thing—electromagnetic waves. (See [link].) What are electromagnetic waves? How are they created, and how do they travel? How can we understand and organize their widely varying properties? What is their relationship to electric and magnetic effects? These and other questions will be explored.

Note:

Misconception Alert: Sound Waves vs. Radio Waves

Many people confuse sound waves with **radio waves**, one type of electromagnetic (EM) wave. However, sound and radio waves are completely different phenomena. Sound creates pressure variations (waves) in matter, such as air or water, or your eardrum. Conversely, radio waves are *electromagnetic waves*, like visible light, infrared, ultraviolet, X-rays, and gamma rays. EM waves don’t need a medium in which to propagate; they can travel through a vacuum, such as outer space.

A radio works because sound waves played by the D.J. at the radio station are converted into electromagnetic waves, then encoded and transmitted in the radio-frequency range. The radio in your car receives the radio waves, decodes the information, and uses a speaker to change it back into a sound wave, bringing sweet music to your ears.

Discovering a New Phenomenon

It is worth noting at the outset that the general phenomenon of electromagnetic waves was predicted by theory before it was realized that light is a form of electromagnetic wave. The prediction was made by James Clerk Maxwell in the mid-19th century when he formulated a single theory combining all the electric and magnetic effects known by scientists at that time. “Electromagnetic waves” was the name he gave to the phenomena his theory predicted.

Such a theoretical prediction followed by experimental verification is an indication of the power of science in general, and physics in particular. The underlying connections and unity of physics allow certain great minds to solve puzzles without having all the pieces. The prediction of electromagnetic waves is one of the most spectacular examples of this power. Certain others, such as the prediction of antimatter, will be discussed in later modules.

The electromagnetic waves sent and received by this 50-foot radar dish antenna at Kennedy Space Center in Florida are not visible, but help track expendable launch vehicles with high-definition imagery. The first use of this C-band radar dish was for the launch of the Atlas V rocket sending the New Horizons probe

toward Pluto.
(credit: NASA)

Maxwell’s Equations: Electromagnetic Waves Predicted and Observed

- Restate Maxwell’s equations.

The Scotsman James Clerk Maxwell (1831–1879) is regarded as the greatest theoretical physicist of the 19th century. (See [link].) Although he died young, Maxwell not only formulated a complete electromagnetic theory, represented by **Maxwell’s equations**, he also developed the kinetic theory of gases and made significant contributions to the understanding of color vision and the nature of Saturn’s rings.

James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic

waves. (credit:
G. J. Stodart)

Maxwell brought together all the work that had been done by brilliant physicists such as Oersted, Coulomb, Gauss, and Faraday, and added his own insights to develop the overarching theory of electromagnetism. Maxwell’s equations are paraphrased here in words because their mathematical statement is beyond the level of this text. However, the equations illustrate how apparently simple mathematical statements can elegantly unite and express a multitude of concepts—why mathematics is the language of science.

Note:
Maxwell’s Equations

1. **Electric field lines** originate on positive charges and terminate on negative charges. The electric field is defined as the force per unit charge on a test charge, and the strength of the force is related to the electric constant ε_0, also known as the permittivity of free space. From Maxwell’s first equation we obtain a special form of Coulomb’s law known as Gauss’s law for electricity.
2. **Magnetic field lines** are continuous, having no beginning or end. No magnetic monopoles are known to exist. The strength of the magnetic force is related to the magnetic constant μ_0, also known as the permeability of free space. This second of Maxwell’s equations is known as Gauss’s law for magnetism.
3. A changing magnetic field induces an electromotive force (emf) and, hence, an electric field. The direction of the emf opposes the change. This third of Maxwell’s equations is Faraday’s law of induction, and includes Lenz’s law.
4. Magnetic fields are generated by moving charges or by changing electric fields. This fourth of Maxwell’s equations encompasses Ampere’s law and adds another source of magnetism—changing electric fields.

Maxwell’s equations encompass the major laws of electricity and magnetism. What is not so apparent is the symmetry that Maxwell introduced in his mathematical framework. Especially important is his addition of the hypothesis that changing electric fields create magnetic fields. This is exactly analogous (and symmetric) to Faraday’s law of induction and had been suspected for some time, but fits beautifully into Maxwell’s equations.

Symmetry is apparent in nature in a wide range of situations. In contemporary research, symmetry plays a major part in the search for sub-atomic particles using massive multinational particle accelerators such as the new Large Hadron Collider at CERN.

Note:

Making Connections: Unification of Forces

Maxwell’s complete and symmetric theory showed that electric and magnetic forces are not separate, but different manifestations of the same thing—the electromagnetic force. This classical unification of forces is one motivation for current attempts to unify the four basic forces in nature—the gravitational, electrical, strong, and weak nuclear forces.

Since changing electric fields create relatively weak magnetic fields, they could not be easily detected at the time of Maxwell’s hypothesis. Maxwell realized, however, that oscillating charges, like those in AC circuits, produce changing electric fields. He predicted that these changing fields would propagate from the source like waves generated on a lake by a jumping fish.

The waves predicted by Maxwell would consist of oscillating electric and magnetic fields—defined to be an electromagnetic wave (EM wave). Electromagnetic waves would be capable of exerting forces on charges great distances from their source, and they might thus be detectable. Maxwell calculated that electromagnetic waves would propagate at a speed given by the equation

Equation:

$$c=\frac{1}{\sqrt{\mu_0\varepsilon_0}}.$$

When the values for μ_0 and ε_0 are entered into the equation for c, we find that

Equation:

$$c=\frac{1}{\sqrt{\left(8.85\times10^{-12}\ \frac{\text{C}^2}{\text{N}\cdot\text{m}^2}\right)\left(4\pi\times10^{-7}\ \frac{\text{T}\cdot\text{m}}{\text{A}}\right)}}=3.00\times10^{8}\ \text{m/s},$$

which is the speed of light. In fact, Maxwell concluded that light is an electromagnetic wave having such wavelengths that it can be detected by the eye.

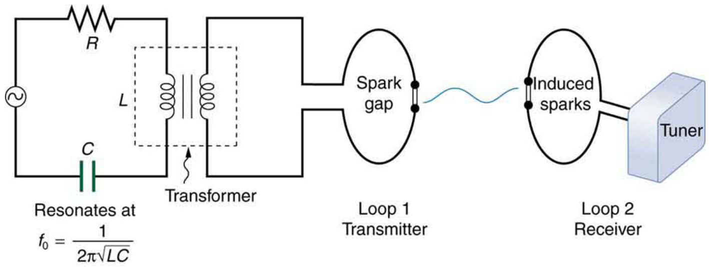
Other wavelengths should exist—it remained to be seen if they did. If so, Maxwell’s theory and remarkable predictions would be verified, the greatest triumph of physics since Newton. Experimental verification came within a few years, but not before Maxwell’s death.

Hertz’s Observations

The German physicist Heinrich Hertz (1857–1894) was the first to generate and detect certain types of electromagnetic waves in the laboratory. Starting in 1887, he performed a series of experiments that not only confirmed the existence of electromagnetic waves, but also verified that they travel at the speed of light.

Hertz used an AC RLC (resistor-inductor-capacitor) circuit that resonates at a known frequency $f_0=\frac{1}{2\pi\sqrt{\text{LC}}}$ and connected it to a loop of wire as shown in [link]. High voltages induced across the gap in the loop produced sparks that were visible evidence of the current in the circuit and that helped generate electromagnetic waves.

Across the laboratory, Hertz had another loop attached to another RLC circuit, which could be tuned (as the dial on a radio) to the same resonant frequency as the first and could, thus, be made to receive electromagnetic waves. This loop also had a gap across which sparks were generated, giving solid evidence that electromagnetic waves had been received.



The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An RLC circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence that the waves had been received.

Hertz also studied the reflection, refraction, and interference patterns of the electromagnetic waves he generated, verifying their wave character. He was able to determine wavelength from the interference patterns, and knowing their frequency, he could calculate the propagation speed using the equation $v = f\lambda$ (velocity—or speed—equals frequency times wavelength). Hertz was thus able to prove that electromagnetic waves travel at the speed of light. The SI unit for frequency, the hertz ($1\ \mathrm{Hz} = 1\ \mathrm{cycle/sec}$), is named in his honor.

Section Summary

- Electromagnetic waves consist of oscillating electric and magnetic fields and propagate at the speed of light c. They were predicted by Maxwell, who also showed that
  Equation:

  $$c=\frac{1}{\sqrt{\mu_0\varepsilon_0}},$$

  where μ_0 is the permeability of free space and ε_0 is the permittivity of free space.
- Maxwell's prediction of electromagnetic waves resulted from his formulation of a complete and symmetric theory of electricity and magnetism, known as Maxwell's equations.
- These four equations are paraphrased in this text, rather than presented numerically, and encompass the major laws of electricity and magnetism. First is Gauss's law for electricity, second is Gauss's law for magnetism, third is Faraday's law of induction, including Lenz's law, and fourth is Ampere's law in a symmetric formulation that adds another source of magnetism—changing electric fields.

Problems & Exercises

Exercise:

Problem:

Verify that the correct value for the speed of light c is obtained when numerical values for the permeability and permittivity of free space (μ_0 and ε_0) are entered into the equation $c=\frac{1}{\sqrt{\mu_0\varepsilon_0}}$.

Exercise:

Problem:

Show that, when SI units for μ_0 and ε_0 are entered, the units given by the right-hand side of the equation in the problem above are m/s.

Glossary

electromagnetic waves
radiation in the form of waves of electric and magnetic energy

Maxwell’s equations
a set of four equations that comprise a complete, overarching theory of electromagnetism

RLC circuit
an electric circuit that includes a resistor, capacitor and inductor

hertz
an SI unit denoting the frequency of an electromagnetic wave, in cycles per second

speed of light
in a vacuum, such as space, the speed of light is a constant 3 x 10^8 m/s

electromotive force (emf)
energy produced per unit charge, drawn from a source that produces an electrical current

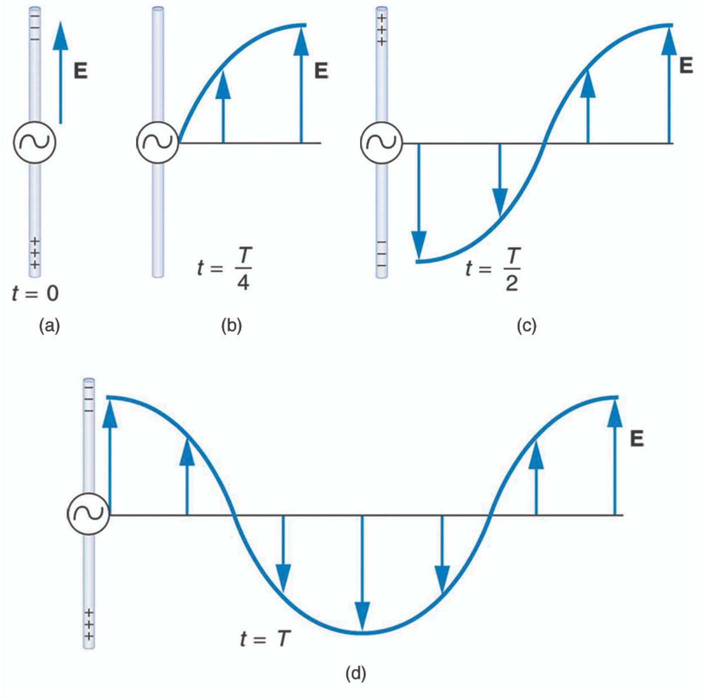
electric field lines
a pattern of imaginary lines that extend between an electric source and charged objects in the surrounding area, with arrows pointed away from positively charged objects and toward negatively charged objects. The more lines in the pattern, the stronger the electric field in that region

magnetic field lines
a pattern of continuous, imaginary lines that emerge from and enter into opposite magnetic poles. The density of the lines indicates the magnitude of the magnetic field

Production of Electromagnetic Waves

- Describe the electric and magnetic waves as they move out from a source, such as an AC generator.
- Explain the mathematical relationship between the magnetic field strength and the electrical field strength.
- Calculate the maximum strength of the magnetic field in an electromagnetic wave, given the maximum electric field strength.

We can get a good understanding of **electromagnetic waves** (EM) by considering how they are produced. Whenever a current varies, associated electric and magnetic fields vary, moving out from the source like waves. Perhaps the easiest situation to visualize is a varying current in a long straight wire, produced by an AC generator at its center, as illustrated in [link].



This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field (**E**) propagates away from the antenna at the speed of light, forming part of an electromagnetic wave.

The **electric field** ($\mathbf{E}$) shown surrounding the wire is produced by the charge distribution on the wire. Both the $\mathbf{E}$ and the charge distribution vary as the current changes. The changing field propagates outward at the speed of light.

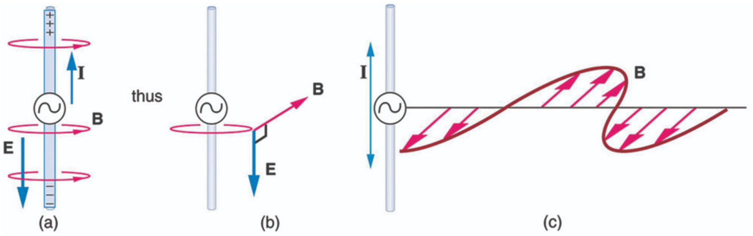
There is an associated **magnetic field** ($\mathbf{B}$) which propagates outward as well (see [link]). The electric and magnetic fields are closely related and propagate as an electromagnetic wave. This is what happens in broadcast antennae such as those in radio and TV stations.

Closer examination of the one complete cycle shown in [link] reveals the periodic nature of the generator-driven charges oscillating up and down in the antenna and the electric field produced. At time $t = 0$, there is the maximum separation of charge, with negative charges at the top and positive charges at the bottom, producing the maximum magnitude of the electric field (or E-field) in the upward direction. One-fourth of a cycle later, there is no charge separation and the field next to the antenna is zero, while the maximum E-field has moved away at speed c.

As the process continues, the charge separation reverses and the field reaches its maximum downward value, returns to zero, and rises to its maximum upward value at the end of one complete cycle. The outgoing wave has an **amplitude** proportional to the maximum separation of charge. Its **wavelength**(λ) is proportional to the period of the oscillation and, hence, is smaller for short periods or high frequencies. (As usual, wavelength and **frequency**(f) are inversely proportional.)

Electric and Magnetic Waves: Moving Together

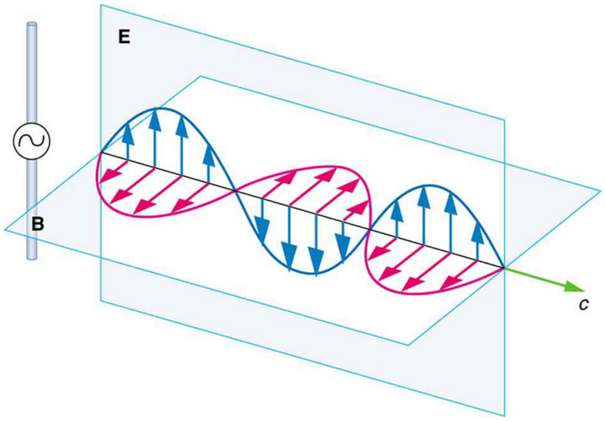
Following Ampere’s law, current in the antenna produces a magnetic field, as shown in [link]. The relationship between $\mathbf{E}$ and $\mathbf{B}$ is shown at one instant in [link] (a). As the current varies, the magnetic field varies in magnitude and direction.



(a) The current in the antenna produces the circular magnetic field lines. The current ($\mathbf{I}$) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields ($\mathbf{E}$ and $\mathbf{B}$) near the wire are perpendicular; they are shown here for one point in space. (c) The magnetic field varies with current and propagates away from the antenna at the speed of light.

The magnetic field lines also propagate away from the antenna at the speed of light, forming the other part of the electromagnetic wave, as seen in [link] (b). The magnetic part of the wave has the same period and wavelength as the electric part, since they are both produced by the same movement and separation of charges in the antenna.

The electric and magnetic waves are shown together at one instant in time in [link]. The electric and magnetic fields produced by a long straight wire antenna are exactly in phase. Note that they are perpendicular to one another and to the direction of propagation, making this a **transverse wave**.



A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields ($\mathbf{E}$ and $\mathbf{B}$) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one direction, but they propagate out in other directions too.

Electromagnetic waves generally propagate out from a source in all directions, sometimes forming a complex radiation pattern. A linear antenna like this one will not radiate parallel to its length, for example. The wave is shown in one direction from the antenna in [link] to illustrate its basic characteristics.

Instead of the AC generator, the antenna can also be driven by an AC circuit. In fact, charges radiate whenever they are accelerated. But while a current in a circuit needs a complete path, an antenna has a varying charge distribution forming a **standing wave**, driven by the AC. The dimensions of the antenna are critical for determining the frequency of the radiated electromagnetic waves. This is a **resonant** phenomenon and when we tune radios or TV, we vary electrical properties to achieve appropriate resonant conditions in the antenna.

Receiving Electromagnetic Waves

Electromagnetic waves carry energy away from their source, similar to a sound wave carrying energy away from a standing wave on a guitar string. An antenna for receiving EM signals works in reverse. And like antennas that produce EM waves, receiver antennas are specially designed to resonate at particular frequencies.

An incoming electromagnetic wave accelerates electrons in the antenna, setting up a standing wave. If the radio or TV is switched on, electrical components pick up and amplify the signal formed by the accelerating electrons. The signal is then converted to audio and/or video format. Sometimes big receiver dishes are used to focus the signal onto an antenna.

In fact, charges radiate whenever they are accelerated. When designing circuits, we often assume that energy does not quickly escape AC circuits, and mostly this is true. A broadcast antenna is specially designed to enhance the rate of electromagnetic radiation, and shielding is necessary to keep the radiation close to zero. Some familiar phenomena are based on the production of electromagnetic waves by varying currents. Your microwave oven, for example, sends electromagnetic waves, called microwaves, from a concealed antenna that has an oscillating current imposed on it.

Relating E-Field and B-Field Strengths

There is a relationship between the E- and B-field strengths in an electromagnetic wave. This can be understood by again considering the antenna just described. The stronger the E-field created by a separation of charge, the greater the current and, hence, the greater the B-field created.

Since current is directly proportional to voltage (Ohm’s law) and voltage is directly proportional to E-field strength, the two should be directly proportional. It can be shown that the magnitudes of the fields do have a constant ratio, equal to the speed of light. That is,

Equation:

$$\frac{E}{B} = c$$

is the ratio of E-field strength to B-field strength in any electromagnetic wave. This is true at all times and at all locations in space. A simple and elegant result.

Example:
Calculating B-Field Strength in an Electromagnetic Wave
What is the maximum strength of the B-field in an electromagnetic wave that has a maximum E-field strength of $1000\ \text{V/m}$?
Strategy
To find the B-field strength, we rearrange the above equation to solve for B, yielding
Equation:

$$B = \frac{E}{c}.$$

Solution
We are given E, and c is the speed of light. Entering these into the expression for B yields
Equation:

$$B = \frac{1000\ \text{V/m}}{3.00 \times 10^{8}\ \text{m/s}} = 3.33 \times 10^{-6}\ \text{T},$$

Where T stands for Tesla, a measure of magnetic field strength.
Discussion
The B-field strength is less than a tenth of the Earth’s admittedly weak magnetic field. This means that a relatively strong electric field of 1000 V/m is accompanied by a relatively weak magnetic field. Note that as this

wave spreads out, say with distance from an antenna, its field strengths become progressively weaker.

The result of this example is consistent with the statement made in the module Maxwell’s Equations: Electromagnetic Waves Predicted and Observed that changing electric fields create relatively weak magnetic fields. They can be detected in electromagnetic waves, however, by taking advantage of the phenomenon of resonance, as Hertz did. A system with the same natural frequency as the electromagnetic wave can be made to oscillate. All radio and TV receivers use this principle to pick up and then amplify weak electromagnetic waves, while rejecting all others not at their resonant frequency.

Note:
Take-Home Experiment: Antennas
For your TV or radio at home, identify the antenna, and sketch its shape. If you don’t have cable, you might have an outdoor or indoor TV antenna. Estimate its size. If the TV signal is between 60 and 216 MHz for basic channels, then what is the wavelength of those EM waves?
Try tuning the radio and note the small range of frequencies at which a reasonable signal for that station is received. (This is easier with digital readout.) If you have a car with a radio and extendable antenna, note the quality of reception as the length of the antenna is changed.

Note:
PhET Explorations: Radio Waves and Electromagnetic Fields
Broadcast radio waves from KPhET. Wiggle the transmitter electron manually or have it oscillate automatically. Display the field as a curve or vectors. The strip chart shows the electron positions at the transmitter and at the receiver.
https://archive.cnx.org/specials/c8dd764c-ae74-11e5-af4c-3375261fa183/radio-waves/#sim-radio-waves

Section Summary

- Electromagnetic waves are created by oscillating charges (which radiate whenever accelerated) and have the same frequency as the oscillation.
- Since the electric and magnetic fields in most electromagnetic waves are perpendicular to the direction in which the wave moves, it is ordinarily a transverse wave.
- The strengths of the electric and magnetic parts of the wave are related by

  Equation:

  $$\frac{E}{B} = \mathrm{c},$$

  which implies that the magnetic field B is very weak relative to the electric field E.

Conceptual Questions

Exercise:

Problem:

The direction of the electric field shown in each part of [link] is that produced by the charge distribution in the wire. Justify the direction shown in each part, using the Coulomb force law and the definition of $\mathbf{E} = \mathbf{F}/q$, where q is a positive test charge.

Exercise:

Problem:

Is the direction of the magnetic field shown in [link] (a) consistent with the right-hand rule for current (RHR-2) in the direction shown in the figure?

Exercise:

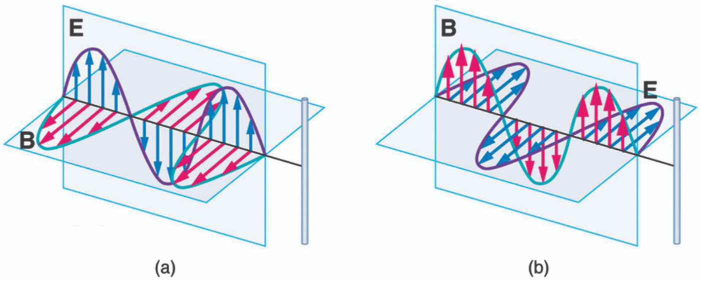
Problem:

Why is the direction of the current shown in each part of [link] opposite to the electric field produced by the wire's charge separation?

Exercise:

Problem:

In which situation shown in [link] will the electromagnetic wave be more successful in inducing a current in the wire? Explain.

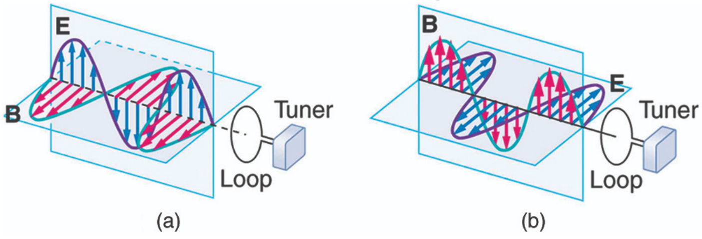


Electromagnetic waves approaching long straight wires.

Exercise:

Problem:

In which situation shown in [link] will the electromagnetic wave be more successful in inducing a current in the loop? Explain.



Electromagnetic waves approaching a wire loop.

Exercise:

Problem:

Should the straight wire antenna of a radio be vertical or horizontal to best receive radio waves broadcast by a vertical transmitter antenna? How should a loop antenna be aligned to best receive the signals? (Note that the direction of the loop that produces the best reception can be used to determine the location of the source. It is used for that purpose in tracking tagged animals in nature studies, for example.)

Exercise:

Problem:

Under what conditions might wires in a DC circuit emit electromagnetic waves?

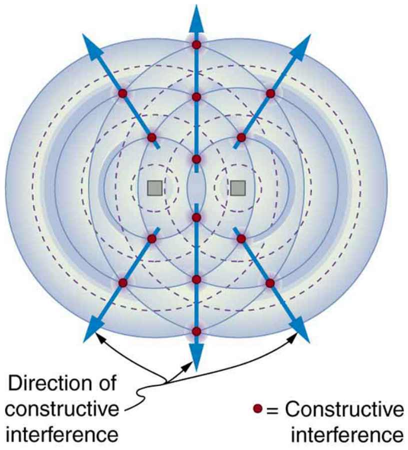
Exercise:

Problem: Give an example of interference of electromagnetic waves.

Exercise:

Problem:

[link] shows the interference pattern of two radio antennas broadcasting the same signal. Explain how this is analogous to the interference pattern for sound produced by two speakers. Could this be used to make a directional antenna system that broadcasts preferentially in certain directions? Explain.


Direction of constructive interference
● = Constructive interference

An overhead view of two radio broadcast antennas sending the same signal, and the interference pattern they produce.

Exercise:

Problem: Can an antenna be any length? Explain your answer.

Problems & Exercises

Exercise:

Problem:

What is the maximum electric field strength in an electromagnetic wave that has a maximum magnetic field strength of $5.00 \times 10^{-4}\ \text{T}$ (about 10 times the Earth’s)?

Solution:

150 kV/m

Exercise:

Problem:

The maximum magnetic field strength of an electromagnetic field is $5 \times 10^{-6}\ \text{T}$. Calculate the maximum electric field strength if the wave is traveling in a medium in which the speed of the wave is $0.75c$.

Exercise:

Problem:

Verify the units obtained for magnetic field strength B in [link] (using the equation $B = \frac{E}{c}$) are in fact teslas (T).

Glossary

electric field
: a vector quantity (**E**); the lines of electric force per unit charge, moving radially outward from a positive charge and in toward a negative charge

electric field strength
: the magnitude of the electric field, denoted *E*-field

magnetic field
: a vector quantity (**B**); can be used to determine the magnetic force on a moving charged particle

magnetic field strength
: the magnitude of the magnetic field, denoted *B*-field

transverse wave
: a wave, such as an electromagnetic wave, which oscillates perpendicular to the axis along the line of travel

standing wave

a wave that oscillates in place, with nodes where no motion happens

wavelength
the distance from one peak to the next in a wave

amplitude
the height, or magnitude, of an electromagnetic wave

frequency
the number of complete wave cycles (up-down-up) passing a given point within one second (cycles/second)

resonant
a system that displays enhanced oscillation when subjected to a periodic disturbance of the same frequency as its natural frequency

oscillate
to fluctuate back and forth in a steady beat

The Electromagnetic Spectrum

- List three “rules of thumb” that apply to the different frequencies along the electromagnetic spectrum.
- Explain why the higher the frequency, the shorter the wavelength of an electromagnetic wave.
- Draw a simplified electromagnetic spectrum, indicating the relative positions, frequencies, and spacing of the different types of radiation bands.
- List and explain the different methods by which electromagnetic waves are produced across the spectrum.

In this module we examine how electromagnetic waves are classified into categories such as radio, infrared, ultraviolet, and so on, so that we can understand some of their similarities as well as some of their differences. We will also find that there are many connections with previously discussed topics, such as wavelength and resonance. A brief overview of the production and utilization of electromagnetic waves is found in [link].

Type of EM wave	**Production**	**Applications**	**Life sciences aspect**	**Issues**
Radio & TV	Accelerating charges	Communications Remote controls	MRI	Requires controls for band use
Microwaves	Accelerating charges & thermal agitation	Communications Ovens Radar	Deep heating	Cell phone use
Infrared	Thermal agitations & electronic transitions	Thermal imaging Heating	Absorbed by atmosphere	Greenhouse effect
Visible light	Thermal agitations & electronic transitions	All pervasive	Photosynthesis Human vision	

Type of EM wave	**Production**	**Applications**	**Life sciences aspect**	**Issues**
Ultraviolet	Thermal agitations & electronic transitions	Sterilization Cancer control	Vitamin D production	Ozone depletion Cancer causing
X-rays	Inner electronic transitions and fast collisions	Medical Security	Medical diagnosis Cancer therapy	Cancer causing
Gamma rays	Nuclear decay	Nuclear medicineSecurity	Medical diagnosis Cancer therapy	Cancer causing Radiation damage

Electromagnetic Waves

Note:
Connections: Waves
There are many types of waves, such as water waves and even earthquakes. Among the many shared attributes of waves are propagation speed, frequency, and wavelength. These are always related by the expression $v_{\text{W}} = f\lambda$. This module concentrates on EM waves, but other modules contain examples of all of these characteristics for sound waves and submicroscopic particles.

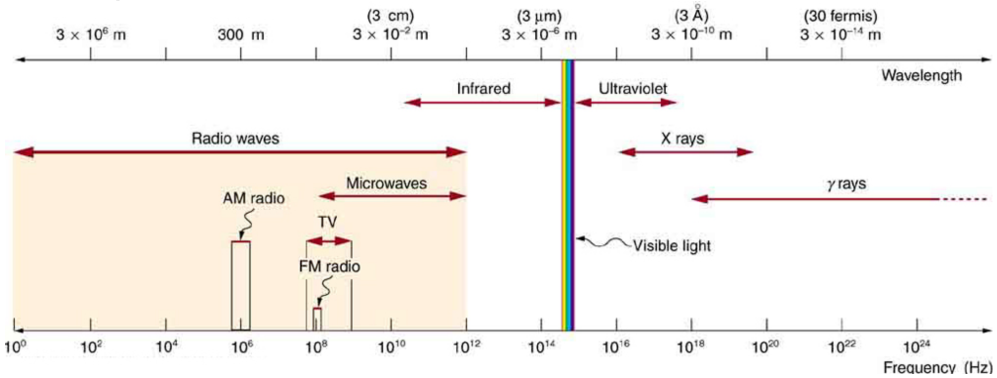
As noted before, an electromagnetic wave has a frequency and a wavelength associated with it and travels at the speed of light, or c. The relationship among these wave characteristics can be described by $v_{\text{W}} = f\lambda$, where v_{W} is the propagation speed of the wave, f is the frequency, and λ is the wavelength. Here $v_{\text{W}} = c$, so that for all electromagnetic waves,
Equation:

$$c = f\lambda.$$

Thus, for all electromagnetic waves, the greater the frequency, the smaller the wavelength.

[link] shows how the various types of electromagnetic waves are categorized according to their wavelengths and frequencies—that is, it shows the electromagnetic spectrum. Many of the

characteristics of the various types of electromagnetic waves are related to their frequencies and wavelengths, as we shall see.



The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

Note:

Electromagnetic Spectrum: Rules of Thumb

Three rules that apply to electromagnetic waves in general are as follows:

- High-frequency electromagnetic waves are more energetic and are more able to penetrate than low-frequency waves.
- High-frequency electromagnetic waves can carry more information per unit time than low-frequency waves.
- The shorter the wavelength of any electromagnetic wave probing a material, the smaller the detail it is possible to resolve.

Note that there are exceptions to these rules of thumb.

Transmission, Reflection, and Absorption

What happens when an electromagnetic wave impinges on a material? If the material is transparent to the particular frequency, then the wave can largely be transmitted. If the material is opaque to the frequency, then the wave can be totally reflected. The wave can also be absorbed by the material, indicating that there is some interaction between the wave and the material, such as the thermal agitation of molecules.

Of course it is possible to have partial transmission, reflection, and absorption. We normally associate these properties with visible light, but they do apply to all electromagnetic waves.

What is not obvious is that something that is transparent to light may be opaque at other frequencies. For example, ordinary glass is transparent to visible light but largely opaque to ultraviolet radiation. Human skin is opaque to visible light—we cannot see through people—but transparent to X-rays.

Radio and TV Waves

The broad category of **radio waves** is defined to contain any electromagnetic wave produced by currents in wires and circuits. Its name derives from their most common use as a carrier of audio information (i.e., radio). The name is applied to electromagnetic waves of similar frequencies regardless of source. Radio waves from outer space, for example, do not come from alien radio stations. They are created by many astronomical phenomena, and their study has revealed much about nature on the largest scales.

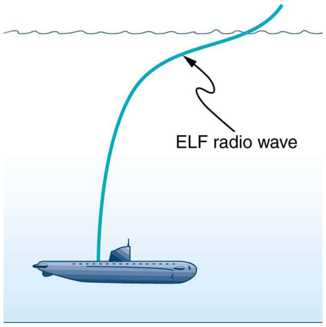
There are many uses for radio waves, and so the category is divided into many subcategories, including microwaves and those electromagnetic waves used for AM and FM radio, cellular telephones, and TV.

The lowest commonly encountered radio frequencies are produced by high-voltage AC power transmission lines at frequencies of 50 or 60 Hz. (See [link].) These extremely long wavelength electromagnetic waves (about 6000 km!) are one means of energy loss in long-distance power transmission.

This high-voltage traction power line running to Eutingen Railway Substation in Germany radiates electromagnetic waves with very long wavelengths. (credit: Zonk43, Wikimedia Commons)

There is an ongoing controversy regarding potential health hazards associated with exposure to these electromagnetic fields (E-fields). Some people suspect that living near such transmission lines may cause a variety of illnesses, including cancer. But demographic data are either inconclusive or simply do not support the hazard theory. Recent reports that have looked at many European and American epidemiological studies have found no increase in risk for cancer due to exposure to E-fields.

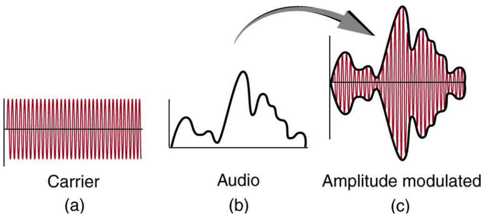
Extremely low frequency (ELF) radio waves of about 1 kHz are used to communicate with submerged submarines. The ability of radio waves to penetrate salt water is related to their wavelength (much like ultrasound penetrating tissue)—the longer the wavelength, the farther they penetrate. Since salt water is a good conductor, radio waves are strongly absorbed by it, and very long wavelengths are needed to reach a submarine under the surface. (See [link].)



Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

AM radio waves are used to carry commercial radio signals in the frequency range from 540 to 1600 kHz. The abbreviation AM stands for **amplitude modulation**, which is the method for placing information on these waves. (See [link].) A **carrier wave** having the basic frequency of the radio station, say 1530 kHz, is varied or modulated in amplitude by an audio signal. The resulting wave has a constant frequency, but a varying amplitude.

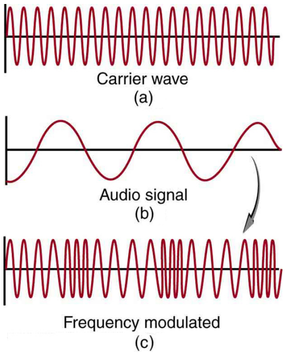
A radio receiver tuned to have the same resonant frequency as the carrier wave can pick up the signal, while rejecting the many other frequencies impinging on its antenna. The receiver’s circuitry is designed to respond to variations in amplitude of the carrier wave to replicate the original audio signal. That audio signal is amplified to drive a speaker or perhaps to be recorded.



Amplitude modulation for AM radio. (a) A carrier wave at the station’s basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

FM Radio Waves

FM radio waves are also used for commercial radio transmission, but in the frequency range of 88 to 108 MHz. FM stands for **frequency modulation**, another method of carrying information. (See [link].) Here a carrier wave having the basic frequency of the radio station, perhaps 105.1 MHz, is modulated in frequency by the audio signal, producing a wave of constant amplitude but varying frequency.



Frequency modulation for

FM radio. (a) A carrier wave at the station’s basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

Since audible frequencies range up to 20 kHz (or 0.020 MHz) at most, the frequency of the FM radio wave can vary from the carrier by as much as 0.020 MHz. Thus the carrier frequencies of two different radio stations cannot be closer than 0.020 MHz. An FM receiver is tuned to resonate at the carrier frequency and has circuitry that responds to variations in frequency, reproducing the audio information.

FM radio is inherently less subject to noise from stray radio sources than AM radio. The reason is that amplitudes of waves add. So an AM receiver would interpret noise added onto the amplitude of its carrier wave as part of the information. An FM receiver can be made to reject amplitudes other than that of the basic carrier wave and only look for variations in frequency. It is thus easier to reject noise from FM, since noise produces a variation in amplitude.

Television is also broadcast on electromagnetic waves. Since the waves must carry a great deal of visual as well as audio information, each channel requires a larger range of frequencies than simple radio transmission. TV channels utilize frequencies in the range of 54 to 88 MHz and 174 to 222 MHz. (The entire FM radio band lies between channels 88 MHz and 174 MHz.) These TV channels are called VHF (for **very high frequency**). Other channels called UHF (for **ultra high frequency**) utilize an even higher frequency range of 470 to 1000 MHz.

The TV video signal is AM, while the TV audio is FM. Note that these frequencies are those of free transmission with the user utilizing an old-fashioned roof antenna. Satellite dishes and cable transmission of TV occurs at significantly higher frequencies and is rapidly evolving with the use of the high-definition or HD format.

Example:
Calculating Wavelengths of Radio Waves

Calculate the wavelengths of a 1530-kHz AM radio signal, a 105.1-MHz FM radio signal, and a 1.90-GHz cell phone signal.

Strategy

The relationship between wavelength and frequency is $c = f\lambda$, where $c = 3.00 \times 10^8 \text{ m/s}$ is the speed of light (the speed of light is only very slightly smaller in air than it is in a vacuum). We can rearrange this equation to find the wavelength for all three frequencies.

Solution

Rearranging gives

Equation:

$$\lambda = \frac{c}{f}.$$

(a) For the $f = 1530 \text{ kHz}$ AM radio signal, then,

Equation:

$$\begin{array}{lll} \lambda & = & \frac{3.00 \times 10^8 \text{ m/s}}{1530 \times 10^3 \text{ cycles/s}} \\ & = & 196 \text{ m.} \end{array}$$

(b) For the $f = 105.1 \text{ MHz}$ FM radio signal,

Equation:

$$\begin{array}{lll} \lambda & = & \frac{3.00 \times 10^8 \text{ m/s}}{105.1 \times 10^6 \text{ cycles/s}} \\ & = & 2.85 \text{ m.} \end{array}$$

(c) And for the $f = 1.90 \text{ GHz}$ cell phone,

Equation:

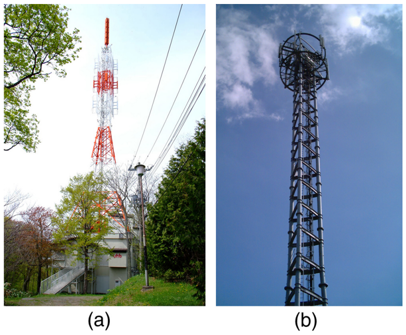
$$\begin{array}{lll} \lambda & = & \frac{3.00 \times 10^8 \text{ m/s}}{1.90 \times 10^9 \text{ cycles/s}} \\ & = & 0.158 \text{ m.} \end{array}$$

Discussion

These wavelengths are consistent with the spectrum in [link]. The wavelengths are also related to other properties of these electromagnetic waves, as we shall see.

The wavelengths found in the preceding example are representative of AM, FM, and cell phones, and account for some of the differences in how they are broadcast and how well they travel. The most efficient length for a linear antenna, such as discussed in Production of Electromagnetic Waves, is $\lambda/2$, half the wavelength of the electromagnetic wave. Thus a very large antenna is needed to efficiently broadcast typical AM radio with its carrier wavelengths on the order of hundreds of meters.

One benefit to these long AM wavelengths is that they can go over and around rather large obstacles (like buildings and hills), just as ocean waves can go around large rocks. FM and TV are best received when there is a line of sight between the broadcast antenna and receiver, and they are often sent from very tall structures. FM, TV, and mobile phone antennas themselves are much smaller than those used for AM, but they are elevated to achieve an unobstructed line of sight. (See [link].)



(a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower—they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa City, Japan. (credit: tokoroten, Wikimedia Commons)

Radio Wave Interference

Astronomers and astrophysicists collect signals from outer space using electromagnetic waves. A common problem for astrophysicists is the “pollution” from electromagnetic radiation pervading our surroundings from communication systems in general. Even everyday gadgets like our car keys having the facility to lock car doors remotely and being able to turn TVs on and off using remotes involve radio-wave frequencies. In order to prevent interference between all these electromagnetic signals, strict regulations are drawn up for different organizations to utilize different radio frequency bands.

One reason why we are sometimes asked to switch off our mobile phones (operating in the range of 1.9 GHz) on airplanes and in hospitals is that important communications or medical equipment often uses similar radio frequencies and their operation can be affected by frequencies used in the communication devices.

For example, radio waves used in magnetic resonance imaging (MRI) have frequencies on the order of 100 MHz, although this varies significantly depending on the strength of the magnetic field used and the nuclear type being scanned. MRI is an important medical imaging and research tool, producing highly detailed two- and three-dimensional images. Radio waves are broadcast, absorbed, and reemitted in a resonance process that is sensitive to the density of nuclei (usually protons or hydrogen nuclei).

The wavelength of 100-MHz radio waves is 3 m, yet using the sensitivity of the resonant frequency to the magnetic field strength, details smaller than a millimeter can be imaged. This is a good example of an exception to a rule of thumb (in this case, the rubric that details much smaller than the probe’s wavelength cannot be detected). The intensity of the radio waves used in MRI presents little or no hazard to human health.

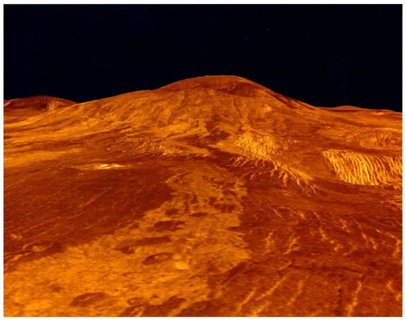
Microwaves

Microwaves are the highest-frequency electromagnetic waves that can be produced by currents in macroscopic circuits and devices. Microwave frequencies range from about 10^{9} Hz to the highest practical LC resonance at nearly 10^{12} Hz. Since they have high frequencies, their wavelengths are short compared with those of other radio waves—hence the name “microwave.”

Microwaves can also be produced by atoms and molecules. They are, for example, a component of electromagnetic radiation generated by **thermal agitation**. The thermal motion of atoms and molecules in any object at a temperature above absolute zero causes them to emit and absorb radiation.

Since it is possible to carry more information per unit time on high frequencies, microwaves are quite suitable for communications. Most satellite-transmitted information is carried on microwaves, as are land-based long-distance transmissions. A clear line of sight between transmitter and receiver is needed because of the short wavelengths involved.

Radar is a common application of microwaves that was first developed in World War II. By detecting and timing microwave echoes, radar systems can determine the distance to objects as diverse as clouds and aircraft. A Doppler shift in the radar echo can be used to determine the speed of a car or the intensity of a rainstorm. Sophisticated radar systems are used to map the Earth and other planets, with a resolution limited by wavelength. (See [link].) The shorter the wavelength of any probe, the smaller the detail it is possible to observe.



An image of Sif Mons with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this image. (credit: NSSDC, NASA/JPL)

Heating with Microwaves

How does the ubiquitous microwave oven produce microwaves electronically, and why does food absorb them preferentially? Microwaves at a frequency of 2.45 GHz are produced by accelerating electrons. The microwaves are then used to induce an alternating electric field in the oven.

Water and some other constituents of food have a slightly negative charge at one end and a slightly positive charge at one end (called polar molecules). The range of microwave frequencies is specially selected so that the polar molecules, in trying to keep orienting themselves with the electric field, absorb these energies and increase their temperatures—called dielectric heating.

The energy thereby absorbed results in thermal agitation heating food and not the plate, which does not contain water. Hot spots in the food are related to constructive and destructive interference patterns. Rotating antennas and food turntables help spread out the hot spots.

Another use of microwaves for heating is within the human body. Microwaves will penetrate more than shorter wavelengths into tissue and so can accomplish “deep heating” (called microwave diathermy). This is used for treating muscular pains, spasms, tendonitis, and rheumatoid arthritis.

Note:
Making Connections: Take-Home Experiment—Microwave Ovens

1. Look at the door of a microwave oven. Describe the structure of the door. Why is there a metal grid on the door? How does the size of the holes in the grid compare with the wavelengths of microwaves used in microwave ovens? What is this wavelength?
2. Place a glass of water (about 250 ml) in the microwave and heat it for 30 seconds. Measure the temperature gain (the ΔT). Assuming that the power output of the oven is 1000 W, calculate the efficiency of the heat-transfer process.
3. Remove the rotating turntable or moving plate and place a cup of water in several places along a line parallel with the opening. Heat for 30 seconds and measure the ΔT for each position. Do you see cases of destructive interference?

Microwaves generated by atoms and molecules far away in time and space can be received and detected by electronic circuits. Deep space acts like a blackbody with a 2.7 K temperature, radiating most of its energy in the microwave frequency range. In 1964, Penzias and Wilson detected this radiation and eventually recognized that it was the radiation of the Big Bang’s cooled remnants.

Infrared Radiation

The microwave and infrared regions of the electromagnetic spectrum overlap (see [link]). **Infrared radiation** is generally produced by thermal motion and the vibration and rotation of atoms and molecules. Electronic transitions in atoms and molecules can also produce infrared radiation.

The range of infrared frequencies extends up to the lower limit of visible light, just below red. In fact, infrared means “below red.” Frequencies at its upper limit are too high to be produced by accelerating electrons in circuits, but small systems, such as atoms and molecules, can vibrate fast enough to produce these waves.

Water molecules rotate and vibrate particularly well at infrared frequencies, emitting and absorbing them so efficiently that the emissivity for skin is $e = 0.97$ in the infrared. Night-vision scopes can detect the infrared emitted by various warm objects, including humans, and convert it to visible light.

We can examine radiant heat transfer from a house by using a camera capable of detecting infrared radiation. Reconnaissance satellites can detect buildings, vehicles, and even individual humans by their infrared emissions, whose power radiation is proportional to the fourth power of the absolute temperature. More mundanely, we use infrared lamps, some of which are called quartz heaters, to preferentially warm us because we absorb infrared better than our surroundings.

The Sun radiates like a nearly perfect blackbody (that is, it has $e = 1$), with a 6000 K surface temperature. About half of the solar energy arriving at the Earth is in the infrared region, with most of the rest in the visible part of the spectrum, and a relatively small amount in the ultraviolet. On average, 50 percent of the incident solar energy is absorbed by the Earth.

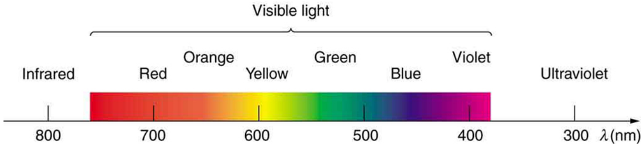
The relatively constant temperature of the Earth is a result of the energy balance between the incoming solar radiation and the energy radiated from the Earth. Most of the infrared radiation emitted from the Earth is absorbed by $\mathrm{CO_2}$ and $\mathrm{H_2O}$ in the atmosphere and then radiated back to Earth or into outer space. This radiation back to Earth is known as the greenhouse effect, and it maintains the surface temperature of the Earth about $40^\mathrm{o}\mathrm{C}$ higher than it would be if there is no absorption. Some scientists think that the increased concentration of $\mathrm{CO_2}$ and other greenhouse gases in the atmosphere, resulting from increases in fossil fuel burning, has increased global average temperatures.

Visible Light

Visible light is the narrow segment of the electromagnetic spectrum to which the normal human eye responds. Visible light is produced by vibrations and rotations of atoms and molecules, as well as by electronic transitions within atoms and molecules. The receivers or detectors of light largely utilize electronic transitions. We say the atoms and molecules are excited when they absorb and relax when they emit through electronic transitions.

[link] shows this part of the spectrum, together with the colors associated with particular pure wavelengths. We usually refer to visible light as having wavelengths of between 400 nm and 750 nm. (The retina of the eye actually responds to the lowest ultraviolet frequencies, but these do not normally reach the retina because they are absorbed by the cornea and lens of the eye.)

Red light has the lowest frequencies and longest wavelengths, while violet has the highest frequencies and shortest wavelengths. Blackbody radiation from the Sun peaks in the visible part of the spectrum but is more intense in the red than in the violet, making the Sun yellowish in appearance.



A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly

distinct, nor are those between the seven
rainbow colors.

Living things—plants and animals—have evolved to utilize and respond to parts of the electromagnetic spectrum they are embedded in. Visible light is the most predominant and we enjoy the beauty of nature through visible light. Plants are more selective. Photosynthesis makes use of parts of the visible spectrum to make sugars.

Example:

Integrated Concept Problem: Correcting Vision with Lasers

During laser vision correction, a brief burst of 193-nm ultraviolet light is projected onto the cornea of a patient. It makes a spot 0.80 mm in diameter and evaporates a layer of cornea $0.30\ \mu\text{m}$ thick. Calculate the energy absorbed, assuming the corneal tissue has the same properties as water; it is initially at 34^{o}C. Assume the evaporated tissue leaves at a temperature of 100^{o}C.

Strategy

The energy from the laser light goes toward raising the temperature of the tissue and also toward evaporating it. Thus we have two amounts of heat to add together. Also, we need to find the mass of corneal tissue involved.

Solution

To figure out the heat required to raise the temperature of the tissue to 100^{o}C, we can apply concepts of thermal energy. We know that

Equation:

$$\text{Q} = \text{mc}\Delta T,$$

where Q is the heat required to raise the temperature, ΔT is the desired change in temperature, m is the mass of tissue to be heated, and c is the specific heat of water equal to 4186 J/kg/K. Without knowing the mass m at this point, we have

Equation:

$$Q = m(4186\ \text{J/kg/K})(100^{\text{o}}\text{C}–34^{\text{o}}\text{C}) = m(276{,}276\ \text{J/kg}) = m(276\ \text{kJ/kg}\ ).$$

The latent heat of vaporization of water is 2256 kJ/kg, so that the energy needed to evaporate mass m is

Equation:

$$Q_{\text{v}} = mL_{\text{v}} = m(2256\ \text{kJ/kg}).$$

To find the mass m, we use the equation $\rho = m/\text{V}$, where ρ is the density of the tissue and V is its volume. For this case,

Equation:

$$\begin{array}{rcl} m & = & \rho \mathrm{V} \\ & = & (1000\ \mathrm{kg/m^3})(\mathrm{area}\times\mathrm{thickness}(\mathrm{m^3})) \\ & = & (1000\ \mathrm{kg/m^3})(\pi(0.80\times10^{-3}\ \mathrm{m})^2/4)(0.30\times10^{-6}\ \mathrm{m}) \\ & = & 0.151\times10^{-9}\ \mathrm{kg}. \end{array}$$

Therefore, the total energy absorbed by the tissue in the eye is the sum of Q and $\mathrm{Q_v}$:
Equation:

$$\mathrm{Q_{tot}} = m(c\Delta\mathrm{T} + \mathrm{L_v}) = (0.151\times10^{-9}\ \mathrm{kg})(276\ \mathrm{kJ/kg} + 2256\ \mathrm{kJ/kg}) = 382\times10^{-9}\ \mathrm{kJ}.$$

Discussion
The lasers used for this eye surgery are excimer lasers, whose light is well absorbed by biological tissue. They evaporate rather than burn the tissue, and can be used for precision work. Most lasers used for this type of eye surgery have an average power rating of about one watt. For our example, if we assume that each laser burst from this pulsed laser lasts for 10 ns, and there are 400 bursts per second, then the average power is $\mathrm{Q_{tot}}\times400 = 150\ \mathrm{mW}.$

Optics is the study of the behavior of visible light and other forms of electromagnetic waves. Optics falls into two distinct categories. When electromagnetic radiation, such as visible light, interacts with objects that are large compared with its wavelength, its motion can be represented by straight lines like rays. Ray optics is the study of such situations and includes lenses and mirrors.

When electromagnetic radiation interacts with objects about the same size as the wavelength or smaller, its wave nature becomes apparent. For example, observable detail is limited by the wavelength, and so visible light can never detect individual atoms, because they are so much smaller than its wavelength. Physical or wave optics is the study of such situations and includes all wave characteristics.

Note:
Take-Home Experiment: Colors That Match
When you light a match you see largely orange light; when you light a gas stove you see blue light. Why are the colors different? What other colors are present in these?

Ultraviolet Radiation

Ultraviolet means “above violet.” The electromagnetic frequencies of **ultraviolet radiation (UV)** extend upward from violet, the highest-frequency visible light. Ultraviolet is also produced by atomic and molecular motions and electronic transitions. The wavelengths of ultraviolet extend from 400 nm down to about 10 nm at its highest frequencies, which overlap with the lowest X-ray frequencies. It was recognized as early as 1801 by Johann Ritter that the solar spectrum had an invisible component beyond the violet range.

Solar UV radiation is broadly subdivided into three regions: UV-A (320–400 nm), UV-B (290–320 nm), and UV-C (220–290 nm), ranked from long to shorter wavelengths (from smaller to larger energies). Most UV-B and all UV-C is absorbed by ozone (O_3) molecules in the upper atmosphere. Consequently, 99% of the solar UV radiation reaching the Earth’s surface is UV-A.

Human Exposure to UV Radiation

It is largely exposure to UV-B that causes skin cancer. It is estimated that as many as 20% of adults will develop skin cancer over the course of their lifetime. Again, treatment is often successful if caught early. Despite very little UV-B reaching the Earth’s surface, there are substantial increases in skin-cancer rates in countries such as Australia, indicating how important it is that UV-B and UV-C continue to be absorbed by the upper atmosphere.

All UV radiation can damage collagen fibers, resulting in an acceleration of the aging process of skin and the formation of wrinkles. Because there is so little UV-B and UV-C reaching the Earth’s surface, sunburn is caused by large exposures, and skin cancer from repeated exposure. Some studies indicate a link between overexposure to the Sun when young and melanoma later in life.

The tanning response is a defense mechanism in which the body produces pigments to absorb future exposures in inert skin layers above living cells. Basically UV-B radiation excites DNA molecules, distorting the DNA helix, leading to mutations and the possible formation of cancerous cells.

Repeated exposure to UV-B may also lead to the formation of cataracts in the eyes—a cause of blindness among people living in the equatorial belt where medical treatment is limited. Cataracts, clouding in the eye’s lens and a loss of vision, are age related; 60% of those between the ages of 65 and 74 will develop cataracts. However, treatment is easy and successful, as one replaces the lens of the eye with a plastic lens. Prevention is important. Eye protection from UV is more effective with plastic sunglasses than those made of glass.

A major acute effect of extreme UV exposure is the suppression of the immune system, both locally and throughout the body.

Low-intensity ultraviolet is used to sterilize haircutting implements, implying that the energy associated with ultraviolet is deposited in a manner different from lower-frequency electromagnetic waves. (Actually this is true for all electromagnetic waves with frequencies greater than visible light.)

Flash photography is generally not allowed of precious artworks and colored prints because the UV radiation from the flash can cause photo-degradation in the artworks. Often artworks will have an extra-thick layer of glass in front of them, which is especially designed to absorb UV radiation.

UV Light and the Ozone Layer

If all of the Sun’s ultraviolet radiation reached the Earth’s surface, there would be extremely grave effects on the biosphere from the severe cell damage it causes. However, the layer of ozone (O_3) in our upper atmosphere (10 to 50 km above the Earth) protects life by absorbing most of the dangerous UV radiation.

Unfortunately, today we are observing a depletion in ozone concentrations in the upper atmosphere. This depletion has led to the formation of an “ozone hole” in the upper atmosphere. The hole is more centered over the southern hemisphere, and changes with the seasons, being largest in the spring. This depletion is attributed to the breakdown of ozone molecules by refrigerant gases called chlorofluorocarbons (CFCs).

The UV radiation helps dissociate the CFC’s, releasing highly reactive chlorine (Cl) atoms, which catalyze the destruction of the ozone layer. For example, the reaction of $\mathrm{CFCl_3}$ with a photon of light (hv) can be written as:

Equation:

$$\mathrm{CFCl_3 + h}v \rightarrow \mathrm{CFCl_2 + Cl.}$$

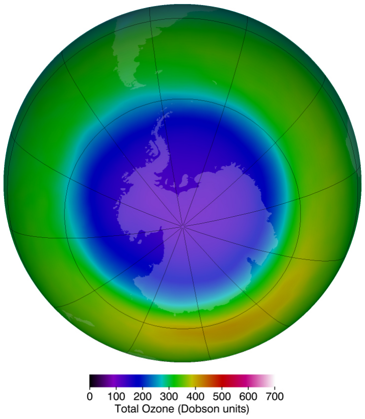
The Cl atom then catalyzes the breakdown of ozone as follows:

Equation:

$$\mathrm{Cl + O_3 \rightarrow ClO + O_2 \text{ and } ClO + O_3 \rightarrow Cl + 2O_2}.$$

A single chlorine atom could destroy ozone molecules for up to two years before being transported down to the surface. The CFCs are relatively stable and will contribute to ozone depletion for years to come. CFCs are found in refrigerants, air conditioning systems, foams, and aerosols.

International concern over this problem led to the establishment of the “Montreal Protocol” agreement (1987) to phase out CFC production in most countries. However, developing-country participation is needed if worldwide production and elimination of CFCs is to be achieved. Probably the largest contributor to CFC emissions today is India. But the protocol seems to be working, as there are signs of an ozone recovery. (See [link].)



This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from the Sun reaches the surface, causing more damage. (credit: NASA Ozone Watch)

Benefits of UV Light

Besides the adverse effects of ultraviolet radiation, there are also benefits of exposure in nature and uses in technology. Vitamin D production in the skin (epidermis) results from exposure to UVB radiation, generally from sunlight. A number of studies indicate lack of vitamin D can result in the development of a range of cancers (prostate, breast, colon), so a certain amount of UV exposure is helpful. Lack of vitamin D is also linked to osteoporosis. Exposures (with no sunscreen) of 10 minutes a day to arms, face, and legs might be sufficient to provide the accepted dietary level. However, in the winter time north of about 37° latitude, most UVB gets blocked by the atmosphere.

UV radiation is used in the treatment of infantile jaundice and in some skin conditions. It is also used in sterilizing workspaces and tools, and killing germs in a wide range of applications. It is also used as an analytical tool to identify substances.

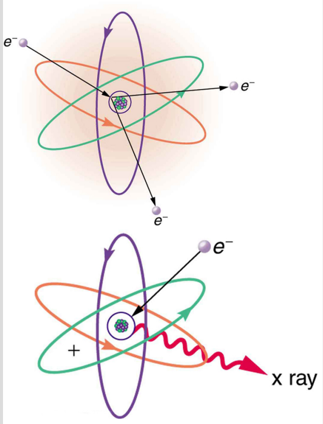
When exposed to ultraviolet, some substances, such as minerals, glow in characteristic visible wavelengths, a process called fluorescence. So-called black lights emit ultraviolet to cause posters and clothing to fluoresce in the visible. Ultraviolet is also used in special microscopes to detect details smaller than those observable with longer-wavelength visible-light microscopes.

Note:

Things Great and Small: A Submicroscopic View of X-Ray Production

X-rays can be created in a high-voltage discharge. They are emitted in the material struck by electrons in the discharge current. There are two mechanisms by which the electrons create X-rays.

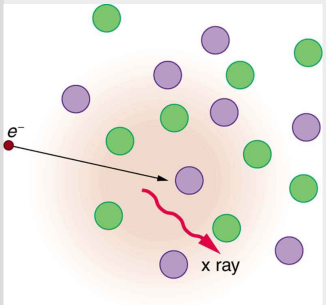
The first method is illustrated in [link]. An electron is accelerated in an evacuated tube by a high positive voltage. The electron strikes a metal plate (e.g., copper) and produces X-rays. Since this is a high-voltage discharge, the electron gains sufficient energy to ionize the atom.



Artist’s conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy released by its fall into a low orbit generates a high-energy EM wave called an X-ray.

In the case shown, an inner-shell electron (one in an orbit relatively close to and tightly bound to the nucleus) is ejected. A short time later, another electron is captured and falls into the orbit in a single great plunge. The energy released by this fall is given to an EM wave known as an X-ray. Since the orbits of the atom are unique to the type of atom, the energy of the X-ray is characteristic of the atom, hence the name characteristic X-ray.

The second method by which an energetic electron creates an X-ray when it strikes a material is illustrated in [link]. The electron interacts with charges in the material as it penetrates. These collisions transfer kinetic energy from the electron to the electrons and atoms in the material.



Artist’s conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are created.

A loss of kinetic energy implies an acceleration, in this case decreasing the electron’s velocity. Whenever a charge is accelerated, it radiates EM waves. Given the high energy of the electron, these EM waves can have high energy. We call them X-rays. Since the process is random, a broad spectrum of X-ray energy is emitted that is more characteristic of the electron energy than the type of material the electron encounters. Such EM radiation is called “bremsstrahlung” (German for “braking radiation”).

X-Rays

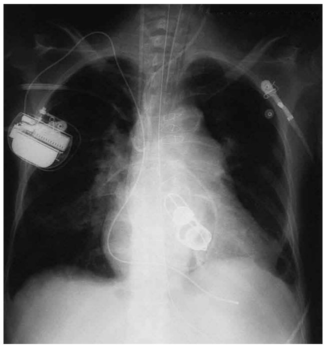
In the 1850s, scientists (such as Faraday) began experimenting with high-voltage electrical discharges in tubes filled with rarefied gases. It was later found that these discharges created an invisible, penetrating form of very high frequency electromagnetic radiation. This radiation was called an **X-ray**, because its identity and nature were unknown.

As described in Things Great and Small, there are two methods by which X-rays are created—both are submicroscopic processes and can be caused by high-voltage discharges. While the low-frequency end of the X-ray range overlaps with the ultraviolet, X-rays extend to much higher frequencies (and energies).

X-rays have adverse effects on living cells similar to those of ultraviolet radiation, and they have the additional liability of being more penetrating, affecting more than the surface layers of cells. Cancer and genetic defects can be induced by exposure to X-rays. Because of their effect on rapidly dividing cells, X-rays can also be used to treat and even cure cancer.

The widest use of X-rays is for imaging objects that are opaque to visible light, such as the human body or aircraft parts. In humans, the risk of cell damage is weighed carefully against the benefit of the diagnostic information obtained. However, questions have risen in recent years as to accidental overexposure of some people during CT scans—a mistake at least in part due to poor monitoring of radiation dose.

The ability of X-rays to penetrate matter depends on density, and so an X-ray image can reveal very detailed density information. [link] shows an example of the simplest type of X-ray image, an X-ray shadow on film. The amount of information in a simple X-ray image is impressive, but more sophisticated techniques, such as CT scans, can reveal three-dimensional information with details smaller than a millimeter.



This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

The use of X-ray technology in medicine is called radiology—an established and relatively cheap tool in comparison to more sophisticated technologies. Consequently, X-rays are widely available and used extensively in medical diagnostics. During World War I, mobile X-ray units, advocated by Madame Marie Curie, were used to diagnose soldiers.

Because they can have wavelengths less than 0.01 nm, X-rays can be scattered (a process called X-ray diffraction) to detect the shape of molecules and the structure of crystals. X-ray diffraction was crucial to Crick, Watson, and Wilkins in the determination of the shape of the double-helix DNA molecule.

X-rays are also used as a precise tool for trace-metal analysis in X-ray induced fluorescence, in which the energy of the X-ray emissions are related to the specific types of elements and amounts of materials present.

Gamma Rays

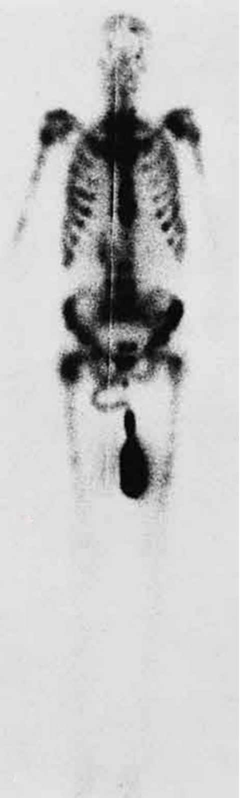
Soon after nuclear radioactivity was first detected in 1896, it was found that at least three distinct types of radiation were being emitted. The most penetrating nuclear radiation was called a **gamma ray (γ ray)** (again a name given because its identity and character were unknown), and it was later found to be an extremely high frequency electromagnetic wave.

In fact, γ rays are any electromagnetic radiation emitted by a nucleus. This can be from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the $\gamma\text{-ray}$ frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation.

Gamma rays have characteristics identical to X-rays of the same frequency—they differ only in source. At higher frequencies, γ rays are more penetrating and more damaging to living tissue. They have many of the same uses as X-rays, including cancer therapy. Gamma radiation from radioactive materials is used in nuclear medicine.

[link] shows a medical image based on γ rays. Food spoilage can be greatly inhibited by exposing it to large doses of γ radiation, thereby obliterating responsible microorganisms. Damage to food cells through irradiation occurs as well, and the long-term hazards of

consuming radiation-preserved food are unknown and controversial for some groups. Both X-ray and γ-ray technologies are also used in scanning luggage at airports.



This is an image of the γ rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

Detecting Electromagnetic Waves from Space

A final note on star gazing. The entire electromagnetic spectrum is used by researchers for investigating stars, space, and time. As noted earlier, Penzias and Wilson detected microwaves to identify the background radiation originating from the Big Bang. Radio telescopes such as the Arecibo Radio Telescope in Puerto Rico and Parkes Observatory in Australia were designed to detect radio waves.

Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

The most famous of these infrared sensitive telescopes is the James Clerk Maxwell Telescope in Hawaii. The earliest telescopes, developed in the seventeenth century, were optical telescopes, collecting visible light. Telescopes in the ultraviolet, X-ray, and γ-ray regions are placed outside the atmosphere on satellites orbiting the Earth.

The Hubble Space Telescope (launched in 1990) gathers ultraviolet radiation as well as visible light. In the X-ray region, there is the Chandra X-ray Observatory (launched in 1999), and in the γ-ray region, there is the new Fermi Gamma-ray Space Telescope (launched in 2008—taking the place of the Compton Gamma Ray Observatory, 1991–2000.).

Note:
PhET Explorations: Color Vision
Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.

Color Vision

Section Summary

- The relationship among the speed of propagation, wavelength, and frequency for any wave is given by $v_{\text{W}} = f\lambda$, so that for electromagnetic waves,
  Equation:
  $$c = f\lambda,$$
  where f is the frequency, λ is the wavelength, and c is the speed of light.
- The electromagnetic spectrum is separated into many categories and subcategories, based on the frequency and wavelength, source, and uses of the electromagnetic waves.
- Any electromagnetic wave produced by currents in wires is classified as a radio wave, the lowest frequency electromagnetic waves. Radio waves are divided into many types, depending on their applications, ranging up to microwaves at their highest frequencies.
- Infrared radiation lies below visible light in frequency and is produced by thermal motion and the vibration and rotation of atoms and molecules. Infrared's lower frequencies overlap with the highest-frequency microwaves.
- Visible light is largely produced by electronic transitions in atoms and molecules, and is defined as being detectable by the human eye. Its colors vary with frequency, from red at the lowest to violet at the highest.
- Ultraviolet radiation starts with frequencies just above violet in the visible range and is produced primarily by electronic transitions in atoms and molecules.
- X-rays are created in high-voltage discharges and by electron bombardment of metal targets. Their lowest frequencies overlap the ultraviolet range but extend to much higher values, overlapping at the high end with gamma rays.
- Gamma rays are nuclear in origin and are defined to include the highest-frequency electromagnetic radiation of any type.

Conceptual Questions

Exercise:

Problem:

If you live in a region that has a particular TV station, you can sometimes pick up some of its audio portion on your FM radio receiver. Explain how this is possible. Does it imply that TV audio is broadcast as FM?

Exercise:

Problem:

Explain why people who have the lens of their eye removed because of cataracts are able to see low-frequency ultraviolet.

Exercise:

Problem:

How do fluorescent soap residues make clothing look “brighter and whiter” in outdoor light? Would this be effective in candlelight?

Exercise:

Problem: Give an example of resonance in the reception of electromagnetic waves.

Exercise:

Problem:

Illustrate that the size of details of an object that can be detected with electromagnetic waves is related to their wavelength, by comparing details observable with two different types (for example, radar and visible light or infrared and X-rays).

Exercise:

Problem: Why don’t buildings block radio waves as completely as they do visible light?

Exercise:

Problem:

Make a list of some everyday objects and decide whether they are transparent or opaque to each of the types of electromagnetic waves.

Exercise:

Problem:

Your friend says that more patterns and colors can be seen on the wings of birds if viewed in ultraviolet light. Would you agree with your friend? Explain your answer.

Exercise:

Problem:

The rate at which information can be transmitted on an electromagnetic wave is proportional to the frequency of the wave. Is this consistent with the fact that laser telephone transmission at visible frequencies carries far more conversations per optical fiber than conventional electronic transmission in a wire? What is the implication for ELF radio communication with submarines?

Exercise:

Problem: Give an example of energy carried by an electromagnetic wave.

Exercise:

Problem:

In an MRI scan, a higher magnetic field requires higher frequency radio waves to resonate with the nuclear type whose density and location is being imaged. What effect does going to a larger magnetic field have on the most efficient antenna to broadcast those radio waves? Does it favor a smaller or larger antenna?

Exercise:

Problem:

Laser vision correction often uses an excimer laser that produces 193-nm electromagnetic radiation. This wavelength is extremely strongly absorbed by the cornea and ablates it in a manner that reshapes the cornea to correct vision defects. Explain how the strong absorption helps concentrate the energy in a thin layer and thus give greater accuracy in shaping the cornea. Also explain how this strong absorption limits damage to the lens and retina of the eye.

Problems & Exercises

Exercise:

Problem:

(a) Two microwave frequencies are authorized for use in microwave ovens: 900 and 2560 MHz. Calculate the wavelength of each. (b) Which frequency would produce smaller hot spots in foods due to interference effects?

Solution:

(a) 33.3 cm (900 MHz) 11.7 cm (2560 MHz)

(b) The microwave oven with the smaller wavelength would produce smaller hot spots in foods, corresponding to the one with the frequency 2560 MHz.

Exercise:

Problem:

(a) Calculate the range of wavelengths for AM radio given its frequency range is 540 to 1600 kHz. (b) Do the same for the FM frequency range of 88.0 to 108 MHz.

Exercise:

Problem:

A radio station utilizes frequencies between commercial AM and FM. What is the frequency of a 11.12-m-wavelength channel?

Solution:

26.96 MHz

Exercise:

Problem:

Find the frequency range of visible light, given that it encompasses wavelengths from 380 to 760 nm.

Exercise:

Problem:

Combing your hair leads to excess electrons on the comb. How fast would you have to move the comb up and down to produce red light?

Solution:

5.0×10^{14} Hz

Exercise:

Problem:

Electromagnetic radiation having a $15.0 - \mu\text{m}$ wavelength is classified as infrared radiation. What is its frequency?

Exercise:

Problem:

Approximately what is the smallest detail observable with a microscope that uses ultraviolet light of frequency 1.20×10^{15} Hz?

Solution:

Equation:

$$\lambda = \frac{c}{f} = \frac{3.00 \times 10^{8}\text{ m/s}}{1.20 \times 10^{15}\text{ Hz}} = 2.50 \times 10^{-7}\text{ m}$$

Exercise:

Problem:

A radar used to detect the presence of aircraft receives a pulse that has reflected off an object 6×10^{-5} s after it was transmitted. What is the distance from the radar station to the reflecting object?

Exercise:

Problem:

Some radar systems detect the size and shape of objects such as aircraft and geological terrain. Approximately what is the smallest observable detail utilizing 500-MHz radar?

Solution:

0.600 m

Exercise:

Problem:

Determine the amount of time it takes for X-rays of frequency $3 \times 10^{18}\text{ Hz}$ to travel (a) 1 mm and (b) 1 cm.

Exercise:

Problem:

If you wish to detect details of the size of atoms (about $1 \times 10^{-10}\text{ m}$) with electromagnetic radiation, it must have a wavelength of about this size. (a) What is its frequency? (b) What type of electromagnetic radiation might this be?

Solution:

(a) $f = \frac{c}{\lambda} = \frac{3.00 \times 10^{8}\text{ m/s}}{1 \times 10^{-10}\text{ m}} = 3 \times 10^{18}\text{ Hz}$

(b) X-rays

Exercise:

Problem:

If the Sun suddenly turned off, we would not know it until its light stopped coming. How long would that be, given that the Sun is $1.50 \times 10^{11}\text{ m}$ away?

Exercise:

Problem:

Distances in space are often quoted in units of light years, the distance light travels in one year. (a) How many meters is a light year? (b) How many meters is it to Andromeda, the nearest large galaxy, given that it is 2.00×10^{6} light years away? (c) The most distant galaxy yet discovered is 12.0×10^{9} light years away. How far is this in meters?

Exercise:

Problem:

A certain 50.0-Hz AC power line radiates an electromagnetic wave having a maximum electric field strength of 13.0 kV/m. (a) What is the wavelength of this very low frequency electromagnetic wave? (b) What is its maximum magnetic field strength?

Solution:

(a) $6.00 \times 10^{6}\ \text{m}$

(b) $4.33 \times 10^{-5}\ \text{T}$

Exercise:

Problem:

During normal beating, the heart creates a maximum 4.00-mV potential across 0.300 m of a person’s chest, creating a 1.00-Hz electromagnetic wave. (a) What is the maximum electric field strength created? (b) What is the corresponding maximum magnetic field strength in the electromagnetic wave? (c) What is the wavelength of the electromagnetic wave?

Exercise:

Problem:

(a) The ideal size (most efficient) for a broadcast antenna with one end on the ground is one-fourth the wavelength ($\lambda/4$) of the electromagnetic radiation being sent out. If a new radio station has such an antenna that is 50.0 m high, what frequency does it broadcast most efficiently? Is this in the AM or FM band? (b) Discuss the analogy of the fundamental resonant mode of an air column closed at one end to the resonance of currents on an antenna that is one-fourth their wavelength.

Solution:

(a) 1.50×10^{6} Hz, AM band
(b) The resonance of currents on an antenna that is 1/4 their wavelength is analogous to the fundamental resonant mode of an air column closed at one end, since the tube also has a length equal to 1/4 the wavelength of the fundamental oscillation.

Exercise:

Problem:

(a) What is the wavelength of 100-MHz radio waves used in an MRI unit? (b) If the frequencies are swept over a ±1.00 range centered on 100 MHz, what is the range of wavelengths broadcast?

Exercise:

Problem:

(a) What is the frequency of the 193-nm ultraviolet radiation used in laser eye surgery? (b) Assuming the accuracy with which this EM radiation can ablate the cornea is directly proportional to wavelength, how much more accurate can this UV be than the shortest visible wavelength of light?

Solution:

(a) $1.55\times 10^{15}\text{ Hz}$

(b) The shortest wavelength of visible light is 380 nm, so that

Equation:

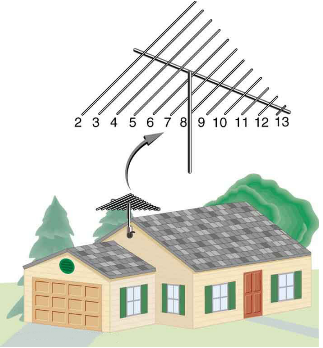
$$\frac{\lambda_{\text{visible}}}{\lambda_{\text{UV}}} = \frac{380\text{ nm}}{193\text{ nm}} = 1.97.$$

In other words, the UV radiation is 97% more accurate than the shortest wavelength of visible light, or almost twice as accurate!

Exercise:

Problem:

TV-reception antennas for VHF are constructed with cross wires supported at their centers, as shown in [link]. The ideal length for the cross wires is one-half the wavelength to be received, with the more expensive antennas having one for each channel. Suppose you measure the lengths of the wires for particular channels and find them to be 1.94 and 0.753 m long, respectively. What are the frequencies for these channels?



A television reception antenna has cross wires of various lengths to most efficiently receive different wavelengths.

Exercise:

Problem:

Conversations with astronauts on lunar walks had an echo that was used to estimate the distance to the Moon. The sound spoken by the person on Earth was transformed into a radio signal sent to the Moon, and transformed back into sound on a speaker inside the astronaut’s space suit. This sound was picked up by the microphone in the space suit (intended for the astronaut’s voice) and sent back to Earth as a radio echo of sorts. If the round-trip time was 2.60 s, what was the approximate distance to the Moon, neglecting any delays in the electronics?

Solution:

$3.90 \times 10^8 \text{ m}$

Exercise:

Problem:

Lunar astronauts placed a reflector on the Moon’s surface, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. (a) To what accuracy in meters can the distance to the Moon be determined, if this time can be measured to 0.100 ns? (b) What percent accuracy is this, given the average distance to the Moon is $3.84 \times 10^8 \text{ m}$?

Exercise:

Problem:

Radar is used to determine distances to various objects by measuring the round-trip time for an echo from the object. (a) How far away is the planet Venus if the echo time is 1000 s? (b) What is the echo time for a car 75.0 m from a Highway Police radar unit? (c) How accurately (in nanoseconds) must you be able to measure the echo time to an airplane 12.0 km away to determine its distance within 10.0 m?

Solution:

(a) $1.50 \times 10^{11} \text{ m}$

(b) $0.500 \ \mu\text{s}$

(c) 66.7 ns

Exercise:

Problem: **Integrated Concepts**

(a) Calculate the ratio of the highest to lowest frequencies of electromagnetic waves the eye can see, given the wavelength range of visible light is from 380 to 760 nm. (b) Compare this with the ratio of highest to lowest frequencies the ear can hear.

Exercise:

Problem: **Integrated Concepts**

(a) Calculate the rate in watts at which heat transfer through radiation occurs (almost entirely in the infrared) from 1.0 m^2 of the Earth's surface at night. Assume the emissivity is 0.90, the temperature of the Earth is 15°C, and that of outer space is 2.7 K. (b) Compare the intensity of this radiation with that coming to the Earth from the Sun during the day, which averages about 800 W/m^2, only half of which is absorbed. (c) What is the maximum magnetic field strength in the outgoing radiation, assuming it is a continuous wave?

Solution:

(a) $-3.5\times10^2 \text{ W/m}^2$

(b) 88%

(c) $1.7 \ \mu\text{T}$

Glossary

electromagnetic spectrum
the full range of wavelengths or frequencies of electromagnetic radiation

radio waves
electromagnetic waves with wavelengths in the range from 1 mm to 100 km; they are produced by currents in wires and circuits and by astronomical phenomena

microwaves
electromagnetic waves with wavelengths in the range from 1 mm to 1 m; they can be produced by currents in macroscopic circuits and devices

thermal agitation
the thermal motion of atoms and molecules in any object at a temperature above absolute zero, which causes them to emit and absorb radiation

radar
a common application of microwaves. Radar can determine the distance to objects as diverse as clouds and aircraft, as well as determine the speed of a car or the intensity of a

rainstorm

infrared radiation (IR)
a region of the electromagnetic spectrum with a frequency range that extends from just below the red region of the visible light spectrum up to the microwave region, or from $0.74\ \mu m$ to $300\ \mu m$

ultraviolet radiation (UV)
electromagnetic radiation in the range extending upward in frequency from violet light and overlapping with the lowest X-ray frequencies, with wavelengths from 400 nm down to about 10 nm

visible light
the narrow segment of the electromagnetic spectrum to which the normal human eye responds

amplitude modulation (AM)
a method for placing information on electromagnetic waves by modulating the amplitude of a carrier wave with an audio signal, resulting in a wave with constant frequency but varying amplitude

extremely low frequency (ELF)
electromagnetic radiation with wavelengths usually in the range of 0 to 300 Hz, but also about 1kHz

carrier wave
an electromagnetic wave that carries a signal by modulation of its amplitude or frequency

frequency modulation (FM)
a method of placing information on electromagnetic waves by modulating the frequency of a carrier wave with an audio signal, producing a wave of constant amplitude but varying frequency

TV
video and audio signals broadcast on electromagnetic waves

very high frequency (VHF)
TV channels utilizing frequencies in the two ranges of 54 to 88 MHz and 174 to 222 MHz

ultra-high frequency (UHF)
TV channels in an even higher frequency range than VHF, of 470 to 1000 MHz

X-ray
invisible, penetrating form of very high frequency electromagnetic radiation, overlapping both the ultraviolet range and the γ-ray range

gamma ray

(γ ray); extremely high frequency electromagnetic radiation emitted by the nucleus of an atom, either from natural nuclear decay or induced nuclear processes in nuclear reactors and weapons. The lower end of the γ-ray frequency range overlaps the upper end of the X-ray range, but γ rays can have the highest frequency of any electromagnetic radiation

Energy in Electromagnetic Waves

- Explain how the energy and amplitude of an electromagnetic wave are related.
- Given its power output and the heating area, calculate the intensity of a microwave oven’s electromagnetic field, as well as its peak electric and magnetic field strengths

Anyone who has used a microwave oven knows there is energy in **electromagnetic waves**. Sometimes this energy is obvious, such as in the warmth of the summer sun. Other times it is subtle, such as the unfelt energy of gamma rays, which can destroy living cells.

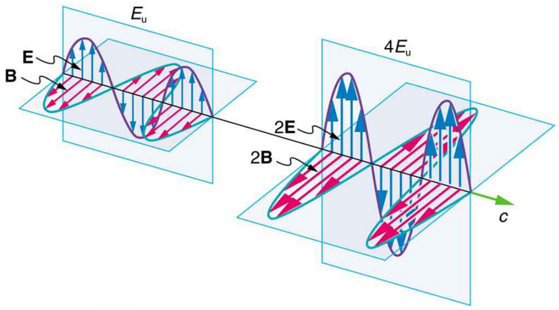
Electromagnetic waves can bring energy into a system by virtue of their **electric and magnetic fields**. These fields can exert forces and move charges in the system and, thus, do work on them. If the frequency of the electromagnetic wave is the same as the natural frequencies of the system (such as microwaves at the resonant frequency of water molecules), the transfer of energy is much more efficient.

Note:

Connections: Waves and Particles

The behavior of electromagnetic radiation clearly exhibits wave characteristics. But we shall find in later modules that at high frequencies, electromagnetic radiation also exhibits particle characteristics. These particle characteristics will be used to explain more of the properties of the electromagnetic spectrum and to introduce the formal study of modern physics.

Another startling discovery of modern physics is that particles, such as electrons and protons, exhibit wave characteristics. This simultaneous sharing of wave and particle properties for all submicroscopic entities is one of the great symmetries in nature.



Energy carried by a wave is proportional to its amplitude squared. With electromagnetic waves, larger E-fields and B-fields exert larger forces and can do more work.

But there is energy in an electromagnetic wave, whether it is absorbed or not. Once created, the fields carry energy away from a source. If absorbed, the field strengths are diminished and anything left travels on. Clearly, the larger the strength of the electric and magnetic fields, the more work they can do and the greater the energy the electromagnetic wave carries.

A wave’s energy is proportional to its **amplitude** squared (E^2 or B^2). This is true for waves on guitar strings, for water waves, and for sound waves, where amplitude is proportional to pressure. In electromagnetic waves, the amplitude is the **maximum field strength** of the electric and magnetic fields. (See [link].)

Thus the energy carried and the **intensity** I of an electromagnetic wave is proportional to E^2 and B^2. In fact, for a continuous sinusoidal electromagnetic wave, the average intensity I_{ave} is given by

Equation:

$$I_{\text{ave}} = \frac{c\varepsilon_0 E_0^2}{2},$$

where c is the speed of light, ε_0 is the permittivity of free space, and E_0 is the maximum electric field strength; intensity, as always, is power per unit area (here in $\mathrm{W/m^2}$).

The average intensity of an electromagnetic wave I_{ave} can also be expressed in terms of the magnetic field strength by using the relationship $B = E/c$, and the fact that $\varepsilon_0 = 1/\mu_0 c^2$, where μ_0 is the permeability of free space. Algebraic manipulation produces the relationship

Equation:

$$I_{\text{ave}} = \frac{\mathrm{cB_0^2}}{2\mathrm{\mu_0}},$$

where B_0 is the maximum magnetic field strength.

One more expression for I_{ave} in terms of both electric and magnetic field strengths is useful. Substituting the fact that $c \cdot B_0 = E_0$, the previous expression becomes

Equation:

$$I_{\text{ave}} = \frac{E_0 B_0}{2\mathrm{\mu_0}}.$$

Whichever of the three preceding equations is most convenient can be used, since they are really just different versions of the same principle: Energy in a wave is related to amplitude squared. Furthermore, since these equations are based on the assumption that the electromagnetic waves are sinusoidal, peak intensity is twice the average; that is, $I_0 = 2I_{\text{ave}}$.

Example:
Calculate Microwave Intensities and Fields
On its highest power setting, a certain microwave oven projects 1.00 kW of microwaves onto a 30.0 by 40.0 cm area. (a) What is the intensity in

$\mathrm{W/m}^2$? (b) Calculate the peak electric field strength E_0 in these waves. (c) What is the peak magnetic field strength B_0?

Strategy

In part (a), we can find intensity from its definition as power per unit area. Once the intensity is known, we can use the equations below to find the field strengths asked for in parts (b) and (c).

Solution for (a)

Entering the given power into the definition of intensity, and noting the area is 0.300 by 0.400 m, yields

Equation:

$$I = \frac{P}{A} = \frac{1.00\ \mathrm{kW}}{0.300\ \mathrm{m}\ \times\ 0.400\ \mathrm{m}}.$$

Here $I = I_{\mathrm{ave}}$, so that

Equation:

$$I_{\mathrm{ave}} = \frac{1000\ \mathrm{W}}{0.120\ \mathrm{m}^2} = 8.33 \times 10^3\ \mathrm{W/m}^2.$$

Note that the peak intensity is twice the average:

Equation:

$$I_0 = 2I_{\mathrm{ave}} = 1.67 \times 10^4\ \mathrm{W/m}^2.$$

Solution for (b)

To find E_0, we can rearrange the first equation given above for I_{ave} to give

Equation:

$$E_0 = \left(\frac{2I_{\mathrm{ave}}}{c\varepsilon_0}\right)^{1/2}.$$

Entering known values gives

Equation:

$$\begin{aligned} E_0 &= \sqrt{\frac{2(8.33\times10^{3}\text{ W/m}^{2})}{(3.00\times10^{8}\text{ m/s})(8.85\times10^{-12}\text{ C}^{2}/\text{N}\cdot\text{m}^{2})}} \\ &= 2.51\times10^{3}\text{ V/m.} \end{aligned}$$

Solution for (c)

Perhaps the easiest way to find magnetic field strength, now that the electric field strength is known, is to use the relationship given by

Equation:

$$B_0 = \frac{E_0}{c}.$$

Entering known values gives

Equation:

$$\begin{aligned} B_0 &= \frac{2.51\times10^{3}\text{ V/m}}{3.0\times10^{8}\text{ m/s}} \\ &= 8.35\times10^{-6}\text{ T.} \end{aligned}$$

Discussion

As before, a relatively strong electric field is accompanied by a relatively weak magnetic field in an electromagnetic wave, since $B = E/c$, and c is a large number.

Section Summary

- The energy carried by any wave is proportional to its amplitude squared. For electromagnetic waves, this means intensity can be expressed as

  Equation:

$$I_{\text{ave}} = \frac{c\varepsilon_0 E_0^2}{2},$$

where I_{ave} is the average intensity in W/m^2, and E_0 is the maximum electric field strength of a continuous sinusoidal wave.

- This can also be expressed in terms of the maximum magnetic field strength B_0 as
  Equation:

$$I_{\text{ave}} = \frac{\text{cB}_0^2}{2\mu_0}$$

  and in terms of both electric and magnetic fields as
  Equation:

$$I_{\text{ave}} = \frac{E_0 B_0}{2\mu_0}.$$

- The three expressions for I_{ave} are all equivalent.

Problems & Exercises

Exercise:

Problem:

What is the intensity of an electromagnetic wave with a peak electric field strength of 125 V/m?

Solution:
Equation:

$$\begin{aligned} I &= \frac{c\varepsilon_0 E_0^2}{2} \\ &= \frac{\left(3.00\times10^8\text{ m/s}\right)\left(8.85\times10^{-12}\text{C}^2/\text{N}\cdot\text{m}^2\right)\left(125\text{ V/m}\right)^2}{2} \\ &= 20.7\text{ W/m}^2 \end{aligned}$$

Exercise:

Problem:

Find the intensity of an electromagnetic wave having a peak magnetic field strength of $4.00\times10^{-9}\text{ T}$.

Exercise:

Problem:

Assume the helium-neon lasers commonly used in student physics laboratories have power outputs of 0.250 mW. (a) If such a laser beam is projected onto a circular spot 1.00 mm in diameter, what is its intensity? (b) Find the peak magnetic field strength. (c) Find the peak electric field strength.

Solution:

(a) $I=\frac{P}{A}=\frac{P}{\pi r^2}=\frac{0.250\times10^{-3}\text{ W}}{\pi\left(0.500\times10^{-3}\text{ m}\right)^2}=318\text{ W/m}^2$

(b)
$$\begin{aligned} I_{\text{ave}} &= \frac{\text{cB}_0^2}{2\text{μ}_0}\Rightarrow B_0=\left(\frac{2\text{μ}_0 I}{c}\right)^{1/2} \\ &= \left(\frac{2\left(4\pi\times10^{-7}\text{ T·m/A}\right)\left(318.3\text{ W/m}^2\right)}{3.00\times10^{8}\text{ m/s}}\right)^{1/2} \\ &= 1.63\times10^{-6}\text{ T} \end{aligned}$$

(c)
$$\begin{aligned} E_0 &= \text{cB}_0=\left(3.00\times10^{8}\text{ m/s}\right)\left(1.633\times10^{-6}\text{ T}\right) \\ &= 4.90\times10^{2}\text{ V/m} \end{aligned}$$

Exercise:

Problem:

An AM radio transmitter broadcasts 50.0 kW of power uniformly in all directions. (a) Assuming all of the radio waves that strike the ground are completely absorbed, and that there is no absorption by the atmosphere or other objects, what is the intensity 30.0 km away? (Hint: Half the power will be spread over the area of a hemisphere.) (b) What is the maximum electric field strength at this distance?

Exercise:

Problem:

Suppose the maximum safe intensity of microwaves for human exposure is taken to be 1.00 W/m^2. (a) If a radar unit leaks 10.0 W of microwaves (other than those sent by its antenna) uniformly in all directions, how far away must you be to be exposed to an intensity considered to be safe? Assume that the power spreads uniformly over the area of a sphere with no complications from absorption or reflection. (b) What is the maximum electric field strength at the safe intensity? (Note that early radar units leaked more than modern ones do. This caused identifiable health problems, such as cataracts, for people who worked near them.)

Solution:

(a) 89.2 cm

(b) 27.4 V/m

Exercise:

Problem:

A 2.50-m-diameter university communications satellite dish receives TV signals that have a maximum electric field strength (for one channel) of $7.50\ \mu\text{V/m}$. (See [link].) (a) What is the intensity of this wave? (b) What is the power received by the antenna? (c) If the orbiting satellite broadcasts uniformly over an area of $1.50\times10^{13}\ \text{m}^2$ (a large fraction of North America), how much power does it radiate?

Satellite dishes receive TV signals sent from orbit. Although the signals are quite weak, the receiver can detect them by being tuned to resonate at their frequency.

Exercise:

Problem:

Lasers can be constructed that produce an extremely high intensity electromagnetic wave for a brief time—called pulsed lasers. They are used to ignite nuclear fusion, for example. Such a laser may produce an electromagnetic wave with a maximum electric field strength of $1.00 \times 10^{11}\ \text{V/m}$ for a time of 1.00 ns. (a) What is the maximum magnetic field strength in the wave? (b) What is the intensity of the beam? (c) What energy does it deliver on a 1.00-mm^2 area?

Solution:

(a) 333 T

(b) $1.33 \times 10^{19}\ \text{W/m}^2$

(c) 13.3 kJ

Exercise:

Problem:

Show that for a continuous sinusoidal electromagnetic wave, the peak intensity is twice the average intensity ($I_0 = 2I_{\text{ave}}$), using either the fact that $E_0 = \sqrt{2}E_{\text{rms}}$, or $B_0 = \sqrt{2}B_{\text{rms}}$, where rms means average (actually root mean square, a type of average).

Exercise:

Problem:

Suppose a source of electromagnetic waves radiates uniformly in all directions in empty space where there are no absorption or interference effects. (a) Show that the intensity is inversely proportional to r^2, the distance from the source squared. (b) Show that the magnitudes of the electric and magnetic fields are inversely proportional to r.

Solution:

(a) $I = \frac{P}{A} = \frac{P}{4\pi r^2} \propto \frac{1}{r^2}$

(b) $I \propto E_0^2, B_0^2 \Rightarrow E_0^2, B_0^2 \propto \frac{1}{r^2} \Rightarrow E_0, B_0 \propto \frac{1}{r}$

Exercise:

Problem: Integrated Concepts

An LC circuit with a 5.00-pF capacitor oscillates in such a manner as to radiate at a wavelength of 3.30 m. (a) What is the resonant frequency? (b) What inductance is in series with the capacitor?

Exercise:

Problem: Integrated Concepts

What capacitance is needed in series with an $800 - \mu\text{H}$ inductor to form a circuit that radiates a wavelength of 196 m?

Solution:

13.5 pF

Exercise:

Problem: Integrated Concepts

Police radar determines the speed of motor vehicles using the same Doppler-shift technique employed for ultrasound in medical diagnostics. Beats are produced by mixing the double Doppler-shifted echo with the original frequency. If 1.50×10^9-Hz microwaves are used and a beat frequency of 150 Hz is produced, what is the speed of the vehicle? (Assume the same Doppler-shift formulas are valid with the speed of sound replaced by the speed of light.)

Exercise:

Problem: Integrated Concepts

Assume the mostly infrared radiation from a heat lamp acts like a continuous wave with wavelength $1.50\ \mu\text{m}$. (a) If the lamp’s 200-W output is focused on a person’s shoulder, over a circular area 25.0 cm in diameter, what is the intensity in W/m^2? (b) What is the peak electric field strength? (c) Find the peak magnetic field strength. (d) How long will it take to increase the temperature of the 4.00-kg shoulder by $2.00^{\text{o}}\ \text{C}$, assuming no other heat transfer and given that its specific heat is $3.47{\times}10^3\ \text{J/kg}{\cdot}^{\text{o}}\text{C}$?

Solution:

(a) $4.07\ \text{kW/m}^2$

(b) 1.75 kV/m

(c) $5.84\ \mu\text{T}$

(d) 2 min 19 s

Exercise:

Problem: Integrated Concepts

On its highest power setting, a microwave oven increases the temperature of 0.400 kg of spaghetti by 45.0^{o}C in 120 s. (a) What was the rate of power absorption by the spaghetti, given that its specific heat is $3.76{\times}10^3\ \text{J/kg}{\cdot}^{\text{o}}\text{C}$? (b) Find the average intensity of the microwaves, given that they are absorbed over a circular area 20.0 cm in diameter. (c) What is the peak electric field strength of the microwave? (d) What is its peak magnetic field strength?

Exercise:

Problem: Integrated Concepts

Electromagnetic radiation from a 5.00-mW laser is concentrated on a 1.00-mm^2 area. (a) What is the intensity in W/m^2? (b) Suppose a 2.00-nC static charge is in the beam. What is the maximum electric force it experiences? (c) If the static charge moves at 400 m/s, what maximum magnetic force can it feel?

Solution:

(a) $5.00 \times 10^3 \text{ W/m}^2$

(b) $3.88 \times 10^{-6} \text{ N}$

(c) $5.18 \times 10^{-12} \text{ N}$

Exercise:

Problem: Integrated Concepts

A 200-turn flat coil of wire 30.0 cm in diameter acts as an antenna for FM radio at a frequency of 100 MHz. The magnetic field of the incoming electromagnetic wave is perpendicular to the coil and has a maximum strength of $1.00 \times 10^{-12} \text{ T}$. (a) What power is incident on the coil? (b) What average emf is induced in the coil over one-fourth of a cycle? (c) If the radio receiver has an inductance of $2.50\ \mu\text{H}$, what capacitance must it have to resonate at 100 MHz?

Exercise:

Problem: Integrated Concepts

If electric and magnetic field strengths vary sinusoidally in time, being zero at $t = 0$, then $E = E_0 \sin 2\pi \text{ft}$ and $B = B_0 \sin 2\pi \text{ft}$. Let $f = 1.00$ GHz here. (a) When are the field strengths first zero? (b) When do they reach their most negative value? (c) How much time is needed for them to complete one cycle?

Solution:

(a) $t = 0$

(b) $7.50 \times 10^{-10} \text{ s}$

(c) 1.00×10^{-9} s

Exercise:

Problem: **Unreasonable Results**

A researcher measures the wavelength of a 1.20-GHz electromagnetic wave to be 0.500 m. (a) Calculate the speed at which this wave propagates. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: **Unreasonable Results**

The peak magnetic field strength in a residential microwave oven is $9.20{\times}10^{-5}\ \text{T}$. (a) What is the intensity of the microwave? (b) What is unreasonable about this result? (c) What is wrong about the premise?

Solution:

(a) $1.01 \times 10^{6}\ \text{W/m}^{2}$

(b) Much too great for an oven.

(c) The assumed magnetic field is unreasonably large.

Exercise:

Problem: **Unreasonable Results**

An LC circuit containing a 2.00-H inductor oscillates at such a frequency that it radiates at a 1.00-m wavelength. (a) What is the capacitance of the circuit? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: **Unreasonable Results**

An LC circuit containing a 1.00-pF capacitor oscillates at such a frequency that it radiates at a 300-nm wavelength. (a) What is the inductance of the circuit? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $2.53 \times 10^{-20}\ \mathrm{H}$

(b) L is much too small.

(c) The wavelength is unreasonably small.

Exercise:

Problem: Create Your Own Problem

Consider electromagnetic fields produced by high voltage power lines. Construct a problem in which you calculate the intensity of this electromagnetic radiation in $\mathrm{W/m^2}$ based on the measured magnetic field strength of the radiation in a home near the power lines. Assume these magnetic field strengths are known to average less than a $\mu\mathrm{T}$. The intensity is small enough that it is difficult to imagine mechanisms for biological damage due to it. Discuss how much energy may be radiating from a section of power line several hundred meters long and compare this to the power likely to be carried by the lines. An idea of how much power this is can be obtained by calculating the approximate current responsible for $\mu\mathrm{T}$ fields at distances of tens of meters.

Exercise:

Problem: Create Your Own Problem

Consider the most recent generation of residential satellite dishes that are a little less than half a meter in diameter. Construct a problem in which you calculate the power received by the dish and the maximum electric field strength of the microwave signals for a single channel received by the dish. Among the things to be considered are the power broadcast by the satellite and the area over which the power is spread, as well as the area of the receiving dish.

Glossary

maximum field strength

the maximum amplitude an electromagnetic wave can reach, representing the maximum amount of electric force and/or magnetic flux that the wave can exert

intensity

the power of an electric or magnetic field per unit area, for example, Watts per square meter

Introduction to Geometric Optics
class="introduction"

Geometric Optics

Light from this page or screen is formed into an image by the lens of your eye, much as the lens of the camera that made this photograph. Mirrors, like lenses, can also form images that in turn are captured by your eye.

Image seen as a result of reflection of light on a plane smooth surface. (credit: NASA Goddard Photo and Video, via Flickr)

Our lives are filled with light. Through vision, the most valued of our senses, light can evoke spiritual emotions, such as when we view a magnificent sunset or glimpse a rainbow breaking through the clouds. Light can also simply amuse us in a theater, or warn us to stop at an intersection. It has innumerable uses beyond vision. Light can carry telephone signals through glass fibers or cook a meal in a solar oven. Life itself could not exist without light’s energy. From photosynthesis in plants to the sun warming a cold-blooded animal, its supply of energy is vital.

Double Rainbow over the bay

of Pocitos in Montevideo, Uruguay. (credit: Madrax, Wikimedia Commons)

We already know that visible light is the type of electromagnetic waves to which our eyes respond. That knowledge still leaves many questions regarding the nature of light and vision. What is color, and how do our eyes detect it? Why do diamonds sparkle? How does light travel? How do lenses and mirrors form images? These are but a few of the questions that are answered by the study of optics. Optics is the branch of physics that deals with the behavior of visible light and other electromagnetic waves. In particular, optics is concerned with the generation and propagation of light and its interaction with matter. What we have already learned about the generation of light in our study of heat transfer by radiation will be expanded upon in later topics, especially those on atomic physics. Now, we will concentrate on the propagation of light and its interaction with matter.

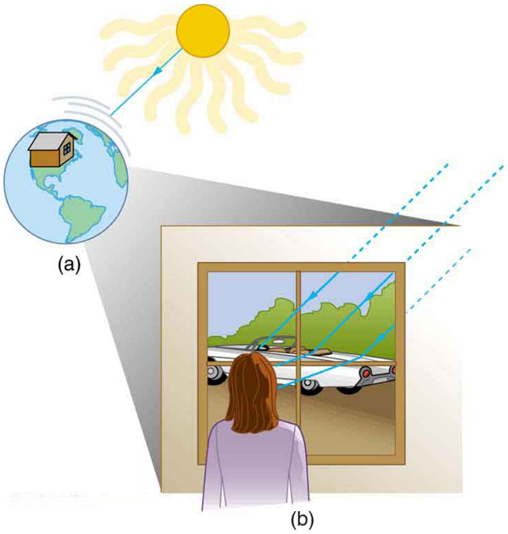
It is convenient to divide optics into two major parts based on the size of objects that light encounters. When light interacts with an object that is several times as large as the light's wavelength, its observable behavior is like that of a ray; it does not prominently display its wave characteristics. We call this part of optics "geometric optics." This chapter will concentrate on such situations. When light interacts with smaller objects, it has very prominent wave characteristics, such as constructive and destructive interference. Wave Optics will concentrate on such situations.

The Ray Aspect of Light

- List the ways by which light travels from a source to another location.

There are three ways in which light can travel from a source to another location. (See [link].) It can come directly from the source through empty space, such as from the Sun to Earth. Or light can travel through various media, such as air and glass, to the person. Light can also arrive after being reflected, such as by a mirror. In all of these cases, light is modeled as traveling in straight lines called rays. Light may change direction when it encounters objects (such as a mirror) or in passing from one material to another (such as in passing from air to glass), but it then continues in a straight line or as a ray. The word **ray** comes from mathematics and here means a straight line that originates at some point. It is acceptable to visualize light rays as laser rays (or even science fiction depictions of ray guns).

Note:
Ray
The word “ray” comes from mathematics and here means a straight line that originates at some point.



Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth traveling through empty space directly from the source. (b) Light can reach a person in one of two ways. It can travel through media like air and glass. It can also reflect from an object like a mirror. In the situations shown here, light interacts with objects large enough that it travels in straight lines, like a ray.

Experiments, as well as our own experiences, show that when light interacts with objects several times as large as its wavelength, it travels in straight lines and acts like a ray. Its wave characteristics are not pronounced in such situations. Since the wavelength of light is less than a micron (a thousandth of a millimeter), it acts like a ray in the many common situations in which it encounters objects larger than a micron. For example, when light encounters anything we can observe with unaided eyes, such as a mirror, it acts like a ray, with only subtle wave characteristics. We will concentrate on the ray characteristics in this chapter.

Since light moves in straight lines, changing directions when it interacts with materials, it is described by geometry and simple trigonometry. This part of optics, where the ray aspect of light dominates, is therefore called **geometric optics**. There are two laws that govern how light changes direction when it interacts with matter. These are the law of reflection, for situations in which light bounces off matter, and the law of refraction, for situations in which light passes through matter.

Note:
Geometric Optics
The part of optics dealing with the ray aspect of light is called geometric optics.

Section Summary

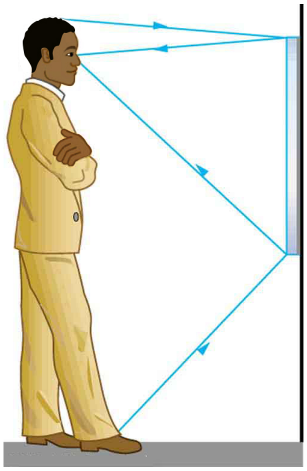
- A straight line that originates at some point is called a ray.
- The part of optics dealing with the ray aspect of light is called geometric optics.
- Light can travel in three ways from a source to another location: (1) directly from the source through empty space; (2) through various media; (3) after being reflected from a mirror.

Problems & Exercises

Exercise:

Problem:

Suppose a man stands in front of a mirror as shown in [link]. His eyes are 1.65 m above the floor, and the top of his head is 0.13 m higher. Find the height above the floor of the top and bottom of the smallest mirror in which he can see both the top of his head and his feet. How is this distance related to the man’s height?



A full-length mirror is one in which you can see all of yourself. It need not be as big as you, and its size is independent of your distance from it.

Solution:

Top from floor, bottom from floor. Height of mirror is , or precisely one-half the height of the person.

Glossary

ray
: straight line that originates at some point

geometric optics

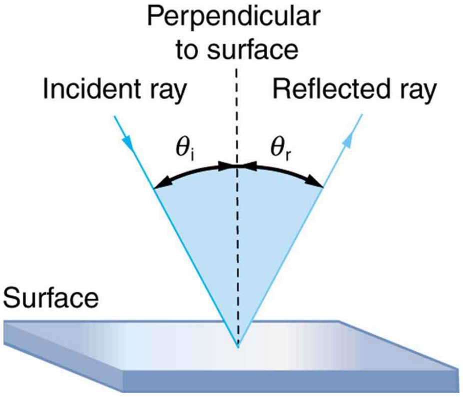
part of optics dealing with the ray aspect of light

The Law of Reflection

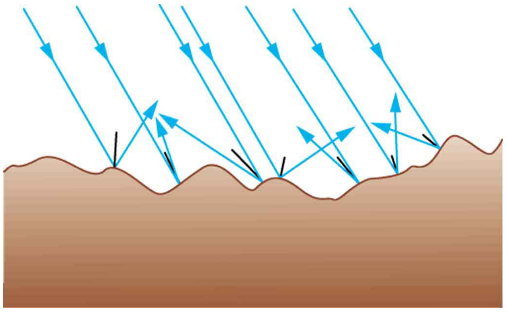
- Explain reflection of light from polished and rough surfaces.

Whenever we look into a mirror, or squint at sunlight glinting from a lake, we are seeing a reflection. When you look at this page, too, you are seeing light reflected from it. Large telescopes use reflection to form an image of stars and other astronomical objects.

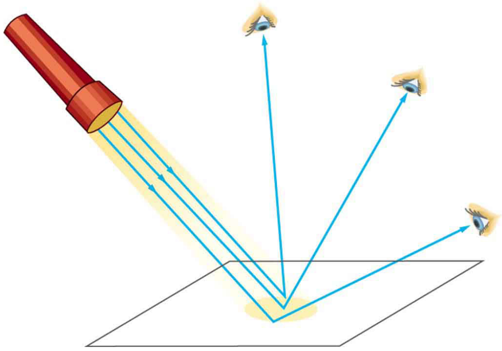
The law of reflection is illustrated in [link], which also shows how the angles are measured relative to the perpendicular to the surface at the point where the light ray strikes. We expect to see reflections from smooth surfaces, but [link] illustrates how a rough surface reflects light. Since the light strikes different parts of the surface at different angles, it is reflected in many different directions, or diffused. Diffused light is what allows us to see a sheet of paper from any angle, as illustrated in [link]. Many objects, such as people, clothing, leaves, and walls, have rough surfaces and can be seen from all sides. A mirror, on the other hand, has a smooth surface (compared with the wavelength of light) and reflects light at specific angles, as illustrated in [link]. When the moon reflects from a lake, as shown in [link], a combination of these effects takes place.



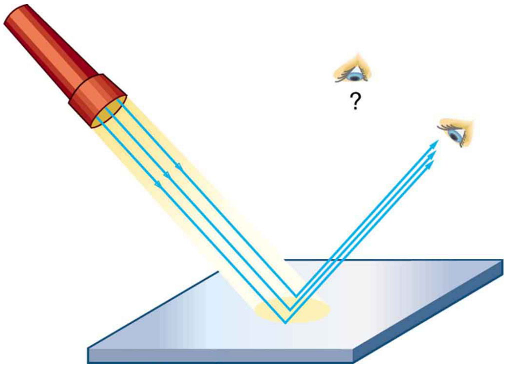
The law of reflection states that the angle of reflection equals the angle of incidence—$\theta_r = \theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.



Light is diffused when it reflects from a rough surface. Here many parallel rays are incident, but they are reflected at many different angles since the surface is rough.



When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light.



A mirror illuminated by many parallel rays reflects them in only one direction, since its surface is very smooth. Only the observer at a particular angle will see the reflected light.

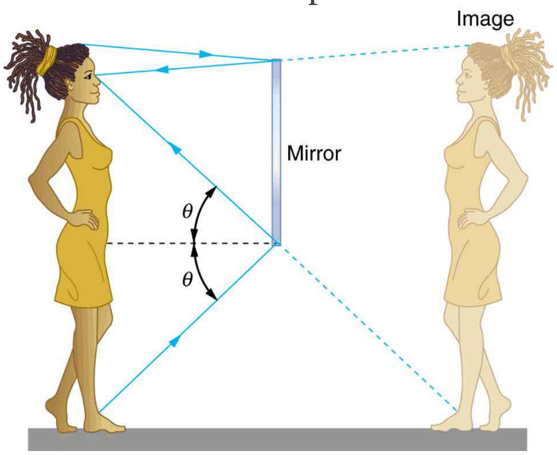
Moonlight is spread out when it is reflected by the lake, since the surface is shiny but uneven. (credit:

Diego Torres Silvestre, Flickr)

The law of reflection is very simple: The angle of reflection equals the angle of incidence.

Note:
The Law of Reflection
The angle of reflection equals the angle of incidence.

When we see ourselves in a mirror, it appears that our image is actually behind the mirror. This is illustrated in [link]. We see the light coming from a direction determined by the law of reflection. The angles are such that our image is exactly the same distance behind the mirror as we stand away from the mirror. If the mirror is on the wall of a room, the images in it are all behind the mirror, which can make the room seem bigger. Although these mirror images make objects appear to be where they cannot be (like behind a solid wall), the images are not figments of our imagination. Mirror images can be photographed and videotaped by instruments and look just as they do with our eyes (optical instruments themselves). The precise manner in which images are formed by mirrors and lenses will be treated in later sections of this chapter.



Our image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be in the direction the rays are coming from when they enter the eyes.

Note:

Take-Home Experiment: Law of Reflection

Take a piece of paper and shine a flashlight at an angle at the paper, as shown in [link]. Now shine the flashlight at a mirror at an angle. Do your observations confirm the predictions in [link] and [link]? Shine the flashlight on various surfaces and determine whether the reflected light is diffuse or not. You can choose a shiny metallic lid of a pot or your skin. Using the mirror and flashlight, can you confirm the law of reflection? You will need to draw lines on a piece of paper showing the incident and reflected rays. (This part works even better if you use a laser pencil.)

Section Summary

- The angle of reflection equals the angle of incidence.
- A mirror has a smooth surface and reflects light at specific angles.
- Light is diffused when it reflects from a rough surface.
- Mirror images can be photographed and videotaped by instruments.

Conceptual Questions

Exercise:

Problem:

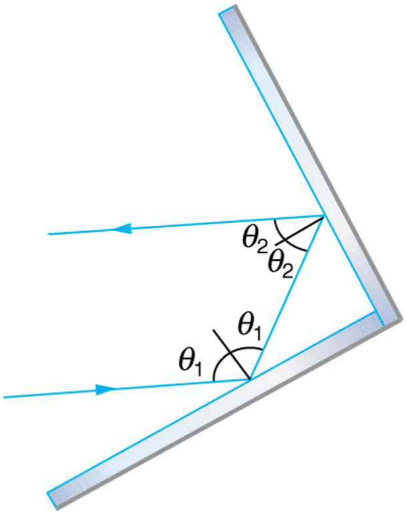
Using the law of reflection, explain how powder takes the shine off of a person’s nose. What is the name of the optical effect?

Problems & Exercises

Exercise:

Problem:

Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated in the following figure.



A corner reflector sends the reflected ray back in a direction parallel to the incident ray, independent of incoming direction.

Exercise:

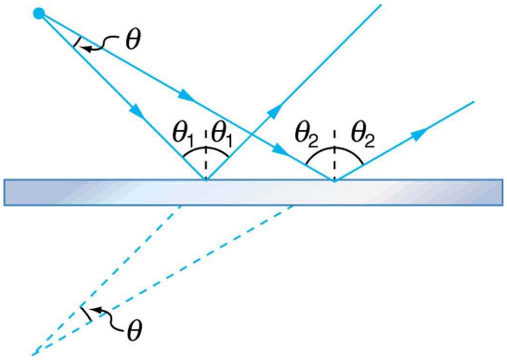
Problem:

Light shows staged with lasers use moving mirrors to swing beams and create colorful effects. Show that a light ray reflected from a mirror changes direction by 2θ when the mirror is rotated by an angle θ.

Exercise:

Problem:

A flat mirror is neither converging nor diverging. To prove this, consider two rays originating from the same point and diverging at an angle θ. Show that after striking a plane mirror, the angle between their directions remains θ.



A flat mirror neither converges nor diverges light rays. Two rays continue to diverge at the same angle after reflection.

Glossary

mirror

smooth surface that reflects light at specific angles, forming an image of the person or object in front of it

law of reflection

angle of reflection equals the angle of incidence

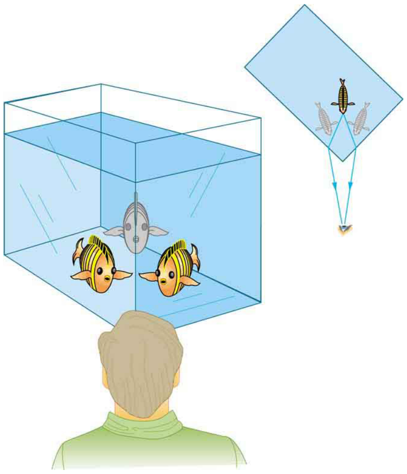
The Law of Refraction

- Determine the index of refraction, given the speed of light in a medium.

It is easy to notice some odd things when looking into a fish tank. For example, you may see the same fish appearing to be in two different places. (See [link].) This is because light coming from the fish to us changes direction when it leaves the tank, and in this case, it can travel two different paths to get to our eyes. The changing of a light ray’s direction (loosely called bending) when it passes through variations in matter is called **refraction**. Refraction is responsible for a tremendous range of optical phenomena, from the action of lenses to voice transmission through optical fibers.

Note:
Refraction
The changing of a light ray’s direction (loosely called bending) when it passes through variations in matter is called refraction.

Note:
Speed of Light
The speed of light *c* not only affects refraction, it is one of the central concepts of Einstein’s theory of relativity. As the accuracy of the measurements of the speed of light were improved, *c* was found not to depend on the velocity of the source or the observer. However, the speed of light does vary in a precise manner with the material it traverses. These facts have far-reaching implications, as we will see in Special Relativity. It makes connections between space and time and alters our expectations that all observers measure the same time for the same event, for example. The speed of light is so important that its value in a vacuum is one of the most fundamental constants in nature as well as being one of the four fundamental SI units.

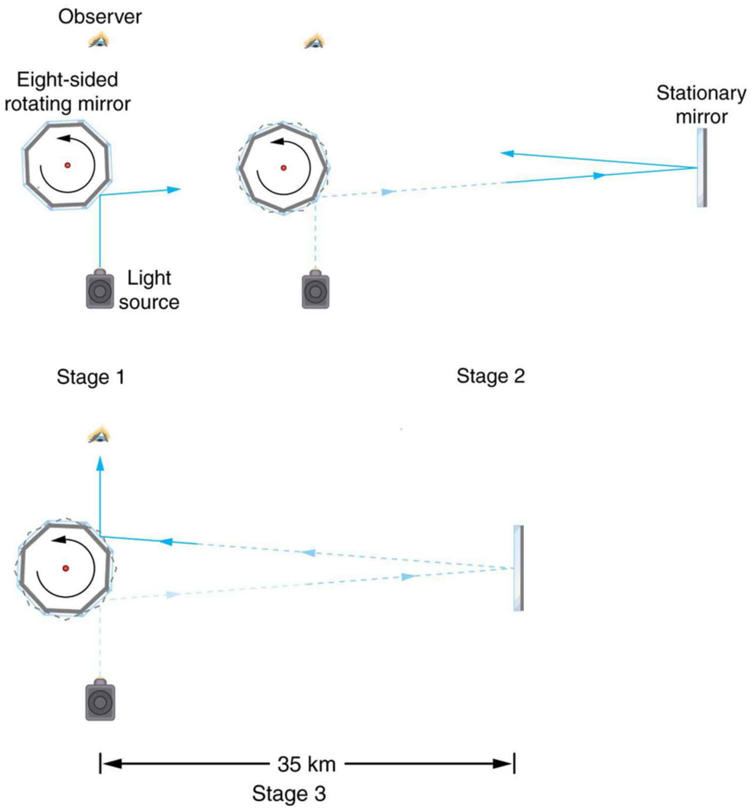


Looking at the fish tank as shown, we can see the same fish in two different locations, because light changes directions when it passes from water to air. In this case, the light can reach the observer by two different paths, and so the fish seems to be in two different places. This bending of light is called refraction and is responsible for many optical phenomena.

Why does light change direction when passing from one material (medium) to another? It is because light changes speed when going from one material to another. So before we study the law of refraction, it is useful to discuss the speed of light and how it varies in different media.

The Speed of Light

Early attempts to measure the speed of light, such as those made by Galileo, determined that light moved extremely fast, perhaps instantaneously. The first real evidence that light traveled at a finite speed came from the Danish astronomer Ole Roemer in the late 17th century. Roemer had noted that the average orbital period of one of Jupiter’s moons, as measured from Earth, varied depending on whether Earth was moving toward or away from Jupiter. He correctly concluded that the apparent change in period was due to the change in distance between Earth and Jupiter and the time it took light to travel this distance. From his 1676 data, a value of the speed of light was calculated to be $2.26 \times 10^8\ \text{m/s}$ (only 25% different from today’s accepted value). In more recent times, physicists have measured the speed of light in numerous ways and with increasing accuracy. One particularly direct method, used in 1887 by the American physicist Albert Michelson (1852–1931), is illustrated in [link]. Light reflected from a rotating set of mirrors was reflected from a stationary mirror 35 km away and returned to the rotating mirrors. The time for the light to travel can be determined by how fast the mirrors must rotate for the light to be returned to the observer’s eye.



A schematic of early apparatus used by Michelson and others to determine the speed of light. As the mirrors rotate, the reflected ray is only briefly directed at the stationary mirror. The returning ray will be reflected into the observer's eye only if the next mirror has rotated into the correct position just as the ray returns. By measuring the correct rotation rate, the time for the round trip can be measured and the speed of light calculated. Michelson’s calculated value of the speed of light was only 0.04% different from the value used today.

The speed of light is now known to great precision. In fact, the speed of light in a vacuum c is so important that it is accepted as one of the basic physical quantities and has the fixed value

Equation:

$$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s},$$

where the approximate value of $3.00 \times 10^8 \text{ m/s}$ is used whenever three-digit accuracy is sufficient. The speed of light through matter is less than it is in a vacuum, because light interacts with atoms in a material. The speed of light depends strongly on the type of material, since its interaction with different atoms, crystal lattices, and other substructures varies. We define the **index of refraction** n of a material to be

Equation:

$$n = \frac{c}{v},$$

where v is the observed speed of light in the material. Since the speed of light is always less than c in matter and equals c only in a vacuum, the index of refraction is always greater than or equal to one.

Note:

Value of the Speed of Light

Equation:

$$c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}$$

Note:

Index of Refraction

Equation:

$$n = \frac{c}{v}$$

That is, $n \geq 1$. [link] gives the indices of refraction for some representative substances. The values are listed for a particular wavelength of light, because they vary slightly with wavelength. (This can have important effects, such as colors produced by a prism.) Note that for gases, n is close to 1.0. This seems reasonable, since atoms in gases are widely separated and light travels at c in the vacuum between atoms. It is common to take $n = 1$ for gases unless great precision is needed. Although the speed of light v in a medium varies considerably from its value c in a vacuum, it is still a large speed.

Medium	***n***
Gases at 0°C***, 1 atm***	
Air	1.000293
Carbon dioxide	1.00045
Hydrogen	1.000139
Oxygen	1.000271
Liquids at 20°C	
Benzene	1.501
Carbon disulfide	1.628

Medium	***n***
Carbon tetrachloride	1.461
Ethanol	1.361
Glycerine	1.473
Water, fresh	1.333
Solids at 20°C	
Diamond	2.419
Fluorite	1.434
Glass, crown	1.52
Glass, flint	1.66
Ice at 20°C	1.309
Polystyrene	1.49
Plexiglas	1.51
Quartz, crystalline	1.544
Quartz, fused	1.458
Sodium chloride	1.544
Zircon	1.923

Index of Refraction in Various Media

Example:
Speed of Light in Matter
Calculate the speed of light in zircon, a material used in jewelry to imitate diamond.

Strategy
The speed of light in a material, v, can be calculated from the index of refraction n of the material using the equation $n = c/v$.

Solution
The equation for index of refraction states that $n = c/v$. Rearranging this to determine v gives

Equation:

$$v = \frac{c}{n}.$$

The index of refraction for zircon is given as 1.923 in [link], and c is given in the equation for speed of light. Entering these values in the last expression gives

Equation:

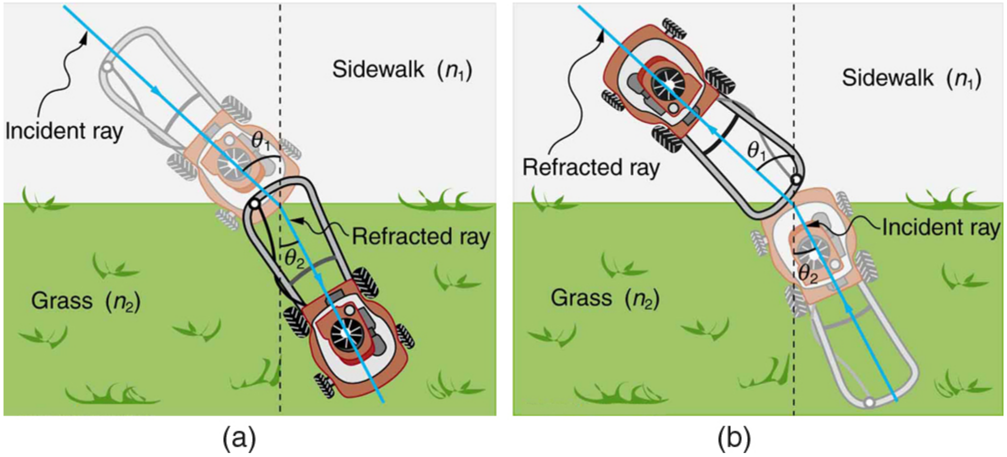
$$\begin{aligned} v &= \frac{3.00 \times 10^8 \text{ m/s}}{1.923} \\ &= 1.56 \times 10^8 \text{ m/s}. \end{aligned}$$

Discussion
This speed is slightly larger than half the speed of light in a vacuum and is still high compared with speeds we normally experience. The only substance listed in [link] that has a greater index of refraction than zircon is diamond. We shall see later that the large index of refraction for zircon makes it sparkle more than glass, but less than diamond.

Law of Refraction

[link] shows how a ray of light changes direction when it passes from one medium to another. As before, the angles are measured relative to a perpendicular to the surface at the point where the light ray crosses it.

(Some of the incident light will be reflected from the surface, but for now we will concentrate on the light that is transmitted.) The change in direction of the light ray depends on how the speed of light changes. The change in the speed of light is related to the indices of refraction of the media involved. In the situations shown in [link], medium 2 has a greater index of refraction than medium 1. This means that the speed of light is less in medium 2 than in medium 1. Note that as shown in [link](a), the direction of the ray moves closer to the perpendicular when it slows down. Conversely, as shown in [link](b), the direction of the ray moves away from the perpendicular when it speeds up. The path is exactly reversible. In both cases, you can imagine what happens by thinking about pushing a lawn mower from a footpath onto grass, and vice versa. Going from the footpath to grass, the front wheels are slowed and pulled to the side as shown. This is the same change in direction as for light when it goes from a fast medium to a slow one. When going from the grass to the footpath, the front wheels can move faster and the mower changes direction as shown. This, too, is the same change in direction as for light going from slow to fast.



The change in direction of a light ray depends on how the speed of light changes when it crosses from one medium to another. The speed of light is greater in medium 1 than in medium 2 in the situations shown here. (a) A ray of light moves closer to the perpendicular when it slows down. This is analogous to what happens when a lawn mower goes from a footpath to grass. (b) A ray of light moves away from the perpendicular when it speeds up. This is analogous to what happens when a lawn mower goes from grass to footpath. The paths are exactly reversible.

The amount that a light ray changes its direction depends both on the incident angle and the amount that the speed changes. For a ray at a given incident angle, a large change in speed causes a large change in direction, and thus a large change in angle. The exact mathematical relationship is the **law of refraction**, or “Snell’s Law,” which is stated in equation form as

Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2.$$

Here n_1 and n_2 are the indices of refraction for medium 1 and 2, and θ_1 and θ_2 are the angles between the rays and the perpendicular in medium 1 and 2, as shown in [link]. The incoming ray is called the incident ray and the outgoing ray the refracted ray, and the associated angles the incident angle and the refracted angle. The law of refraction is also called Snell’s law after the Dutch mathematician Willebrord Snell (1591–1626), who discovered it in 1621. Snell’s experiments showed that the law of refraction was obeyed and that a characteristic index of refraction n could be assigned to a given medium. Snell was not aware that the speed of light varied in different media, but through experiments he was able to determine indices of refraction from the way light rays changed direction.

Note:

The Law of Refraction

Equation:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

Note:

Take-Home Experiment: A Broken Pencil

A classic observation of refraction occurs when a pencil is placed in a glass half filled with water. Do this and observe the shape of the pencil when you look at the pencil sideways, that is, through air, glass, water. Explain your observations. Draw ray diagrams for the situation.

Example:

Determine the Index of Refraction from Refraction Data

Find the index of refraction for medium 2 in [link](a), assuming medium 1 is air and given the incident angle is 30.0º and the angle of refraction is 22.0º.

Strategy

The index of refraction for air is taken to be 1 in most cases (and up to four significant figures, it is 1.000). Thus $n_1 = 1.00$ here. From the given information, $\theta_1 = 30.0^\text{º}$ and $\theta_2 = 22.0^\text{º}$. With this information, the only unknown in Snell's law is n_2, so that it can be used to find this unknown.

Solution

Snell's law is

Equation:

$$n_1 \text{ sin } \theta_1 = n_2 \text{ sin } \theta_2.$$

Rearranging to isolate n_2 gives

Equation:

$$n_2 = n_1 \frac{\text{sin } \theta_1}{\text{sin } \theta_2}.$$

Entering known values,

Equation:

$$\begin{array}{lll} n_2 & = & 1.00 \frac{\text{sin } 30.0^\text{º}}{\text{sin } 22.0^\text{º}} = \frac{0.500}{0.375} \\ & = & 1.33. \end{array}$$

Discussion

This is the index of refraction for water, and Snell could have determined it by measuring the angles and performing this calculation. He would then have found 1.33 to be the appropriate index of refraction for water in all other situations, such as when a ray passes from water to glass. Today we can verify that the index of refraction is related to the speed of light in a medium by measuring that speed directly.

Example:

A Larger Change in Direction

Suppose that in a situation like that in [link], light goes from air to diamond and that the incident angle is 30.0º. Calculate the angle of refraction θ_2 in the diamond.

Strategy

Again the index of refraction for air is taken to be $n_1 = 1.00$, and we are given $\theta_1 = 30.0º$. We can look up the index of refraction for diamond in [link], finding $n_2 = 2.419$. The only unknown in Snell's law is θ_2, which we wish to determine.

Solution

Solving Snell's law for sin θ_2 yields

Equation:

$$\text{sin }\theta_2 = \frac{n_1}{n_2}\text{sin }\theta_1.$$

Entering known values,

Equation:

$$\text{sin }\theta_2 = \frac{1.00}{2.419}\text{sin }30.0º{=}\left(0.413\right)(0.500) = 0.207.$$

The angle is thus

Equation:

$$\theta_2 = \text{sin}^{-1}0.207 = 11.9º.$$

Discussion

For the same 30^o angle of incidence, the angle of refraction in diamond is significantly smaller than in water (11.9^o rather than 22^o—see the preceding example). This means there is a larger change in direction in diamond. The cause of a large change in direction is a large change in the index of refraction (or speed). In general, the larger the change in speed, the greater the effect on the direction of the ray.

Section Summary

- The changing of a light ray’s direction when it passes through variations in matter is called refraction.
- The speed of light in vacuum $c = 2.99792458 \times 10^8 \text{ m/s} \approx 3.00 \times 10^8 \text{ m/s}.$
- Index of refraction $n = \frac{c}{v}$, where v is the speed of light in the material, c is the speed of light in vacuum, and n is the index of refraction.
- Snell’s law, the law of refraction, is stated in equation form as $n_1 \sin \theta_1 = n_2 \sin \theta_2$.

Conceptual Questions

Exercise:

Problem:

Diffusion by reflection from a rough surface is described in this chapter. Light can also be diffused by refraction. Describe how this occurs in a specific situation, such as light interacting with crushed ice.

Exercise:

Problem:

Why is the index of refraction always greater than or equal to 1?

Exercise:

Problem:

Does the fact that the light flash from lightning reaches you before its sound prove that the speed of light is extremely large or simply that it is greater than the speed of sound? Discuss how you could use this effect to get an estimate of the speed of light.

Exercise:

Problem:

Will light change direction toward or away from the perpendicular when it goes from air to water? Water to glass? Glass to air?

Exercise:

Problem:

Explain why an object in water always appears to be at a depth shallower than it actually is? Why do people sometimes sustain neck and spinal injuries when diving into unfamiliar ponds or waters?

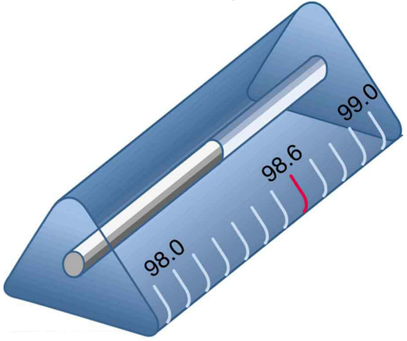
Exercise:

Problem:

Explain why a person’s legs appear very short when wading in a pool. Justify your explanation with a ray diagram showing the path of rays from the feet to the eye of an observer who is out of the water.

Exercise:

Problem: Why is the front surface of a thermometer curved as shown?



The curved surface of the thermometer serves a purpose.

Exercise:

Problem:

Suppose light were incident from air onto a material that had a negative index of refraction, say –1.3; where does the refracted light ray go?

Problems & Exercises

Exercise:

Problem: What is the speed of light in water? In glycerine?

Solution:

$2.25 \times 10^8 \text{ m/s}$ in water

$2.04 \times 10^8 \text{ m/s}$ in glycerine

Exercise:

Problem: What is the speed of light in air? In crown glass?

Exercise:

Problem:

Calculate the index of refraction for a medium in which the speed of light is $2.012 \times 10^8 \text{ m/s}$, and identify the most likely substance based on [link].

Solution:

1.490, polystyrene

Exercise:

Problem:

In what substance in [link] is the speed of light $2.290 \times 10^8 \text{ m/s}$?

Exercise:

Problem:

There was a major collision of an asteroid with the Moon in medieval times. It was described by monks at Canterbury Cathedral in England as a red glow on and around the Moon. How long after the asteroid hit the Moon, which is $3.84 \times 10^5 \text{ km}$ away, would the light first arrive on Earth?

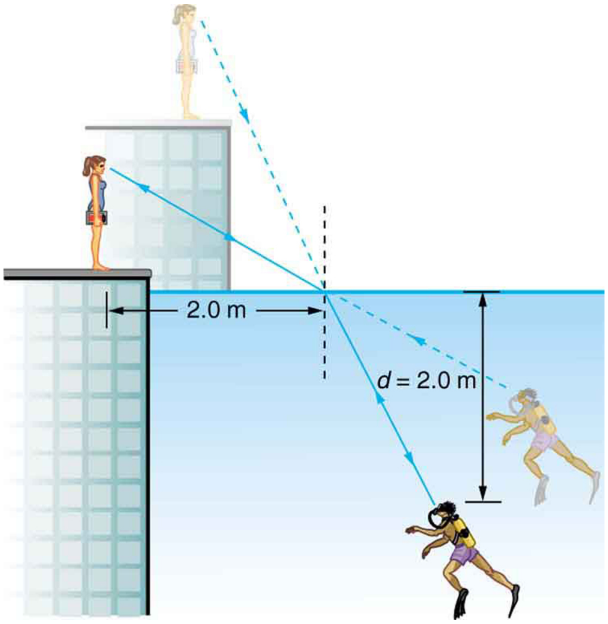
Solution:

1.28 s

Exercise:

Problem:

A scuba diver training in a pool looks at his instructor as shown in [link]. What angle does the ray from the instructor's face make with the perpendicular to the water at the point where the ray enters? The angle between the ray in the water and the perpendicular to the water is 25.0^{o}.



A scuba diver in a pool and his trainer look at each other.

Exercise:

Problem:

Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?

Solution:

1.03 ns

Exercise:

Problem:

(a) Given that the angle between the ray in the water and the perpendicular to the water is 25.0^{o}, and using information in [link], find the height of the instructor’s head above the water, noting that you will first have to calculate the angle of incidence. (b) Find the apparent depth of the diver’s head below water as seen by the instructor.

Exercise:

Problem:

Suppose you have an unknown clear substance immersed in water, and you wish to identify it by finding its index of refraction. You arrange to have a beam of light enter it at an angle of 45.0^{o}, and you observe the angle of refraction to be 40.3^{o}. What is the index of refraction of the substance and its likely identity?

Solution:

$n = 1.46$, fused quartz

Exercise:

Problem:

On the Moon’s surface, lunar astronauts placed a corner reflector, off which a laser beam is periodically reflected. The distance to the Moon is calculated from the round-trip time. What percent correction is needed to account for the delay in time due to the slowing of light in Earth’s atmosphere? Assume the distance to the Moon is precisely $3.84 \times 10^{8}\ \text{m}$, and Earth’s atmosphere (which varies in density with altitude) is equivalent to a layer 30.0 km thick with a constant index of refraction $n = 1.000293$.

Exercise:

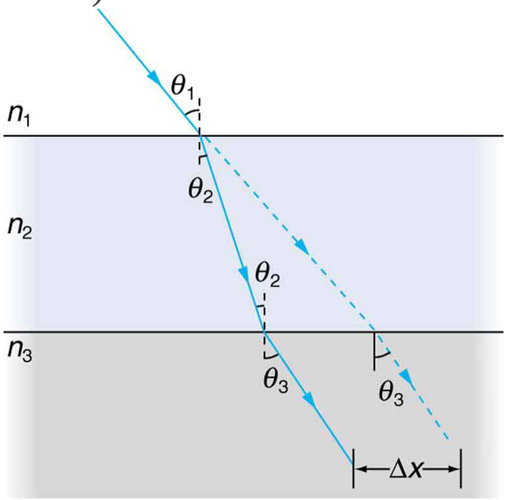
Problem:

Suppose [link] represents a ray of light going from air through crown glass into water, such as going into a fish tank. Calculate the amount the ray is displaced by the glass (Δx), given that the incident angle is 40.0º and the glass is 1.00 cm thick.

Exercise:

Problem:

[link] shows a ray of light passing from one medium into a second and then a third. Show that θ_3 is the same as it would be if the second medium were not present (provided total internal reflection does not occur).



A ray of light passes from one medium to a third by traveling through a second. The final direction is the same as if the second medium were not present, but the ray is displaced by Δx (shown exaggerated).

Exercise:

Problem: **Unreasonable Results**

Suppose light travels from water to another substance, with an angle of incidence of 10.0^{o} and an angle of refraction of 14.9^{o}. (a) What is the index of refraction of the other substance? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) 0.898

(b) Can't have $n < 1.00$ since this would imply a speed greater than c.

(c) Refracted angle is too big relative to the angle of incidence.

Exercise:

Problem: **Construct Your Own Problem**

Consider sunlight entering the Earth's atmosphere at sunrise and sunset—that is, at a 90^{o} incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, both at sunrise and sunset. Now construct a problem in which you determine the angle of refraction for different models of the atmosphere, such as various layers of varying density. Your instructor may wish to guide you on the level of complexity to consider and on how the index of refraction varies with air density.

Exercise:

Problem: **Unreasonable Results**

Light traveling from water to a gemstone strikes the surface at an angle of 80.0^{o} and has an angle of refraction of 15.2^{o}. (a) What is the speed of light in the gemstone? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $\frac{c}{5.00}$

(b) Speed of light too slow, since index is much greater than that of diamond.

(c) Angle of refraction is unreasonable relative to the angle of incidence.

Glossary

refraction
: changing of a light ray’s direction when it passes through variations in matter

index of refraction
: for a material, the ratio of the speed of light in vacuum to that in the material

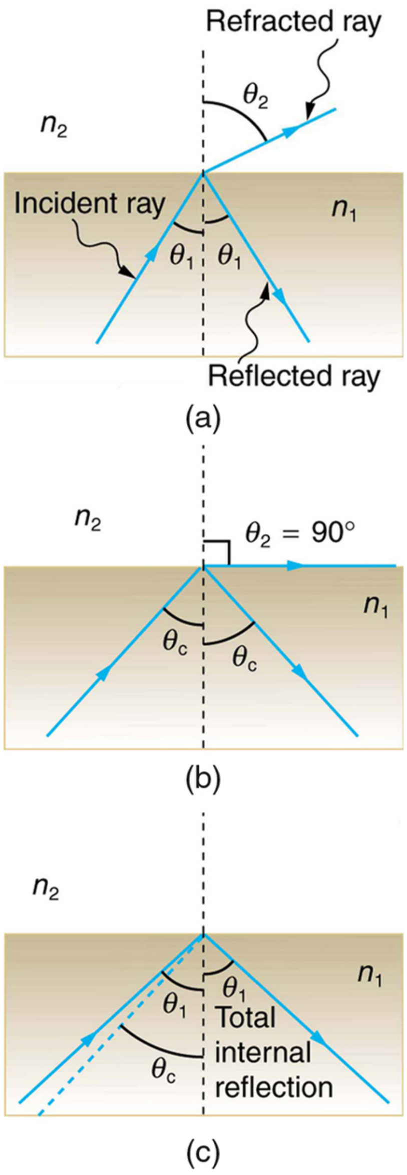
Total Internal Reflection

- Explain the phenomenon of total internal reflection.
- Describe the workings and uses of fiber optics.
- Analyze the reason for the sparkle of diamonds.

A good-quality mirror may reflect more than 90% of the light that falls on it, absorbing the rest. But it would be useful to have a mirror that reflects all of the light that falls on it. Interestingly, we can produce *total reflection* using an aspect of *refraction*.

Consider what happens when a ray of light strikes the surface between two materials, such as is shown in [link](a). Part of the light crosses the boundary and is refracted; the rest is reflected. If, as shown in the figure, the index of refraction for the second medium is less than for the first, the ray bends away from the perpendicular. (Since $n_1 > n_2$, the angle of refraction is greater than the angle of incidence—that is, $\theta_2 > \theta_1$.) Now imagine what happens as the incident angle is increased. This causes θ_2 to increase also. The largest the angle of refraction $\theta_2$can be is 90º, as shown in [link](b).The **critical angle**θ_c for a combination of materials is defined to be the incident angle θ_1 that produces an angle of refraction of 90º. That is, θ_c is the incident angle for which $\theta_2 = 90º$. If the incident angle θ_1 is greater than the critical angle, as shown in [link](c), then all of the light is reflected back into medium 1, a condition called **total internal reflection**.

Note:
Critical Angle
The incident angle θ_1 that produces an angle of refraction of 90º is called the critical angle, θ_c.



(a) A ray of light crosses a boundary where the speed of light increases and the index of refraction decreases. That is, $n_2 < n_1$. The ray bends away from the perpendicular. (b) The critical angle θ_c is the one for which the angle of refraction is . (c) Total internal reflection occurs when the incident angle is greater than the critical angle.

Snell’s law states the relationship between angles and indices of refraction. It is given by

Equation:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

When the incident angle equals the critical angle ($\theta_1 = \theta_c$), the angle of refraction is 90º ($\theta_2 = 90º$). Noting that sin 90º=1, Snell’s law in this case becomes

Equation:

$$n_1 \sin\theta_1 = n_2.$$

The critical angle θ_c for a given combination of materials is thus

Equation:

$$\theta_c = \sin^{-1}(n_2/n_1) \text{ for } n_1 > n_2.$$

Total internal reflection occurs for any incident angle greater than the critical angle θ_c, and it can only occur when the second medium has an index of refraction less than the first. Note the above equation is written for a light ray that travels in medium 1 and reflects from medium 2, as shown in the figure.

Example:

How Big is the Critical Angle Here?

What is the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air?

Strategy

The index of refraction for polystyrene is found to be 1.49 in [link], and the index of refraction of air can be taken to be 1.00, as before. Thus, the condition that the second medium (air) has an index of refraction less than the first (plastic) is satisfied, and the equation $\theta_c = \sin^{-1}(n_2/n_1)$ can be used to find the critical angle θ_c. Here, then, $n_2 = 1.00$ and $n_1 = 1.49$.

Solution

The critical angle is given by

Equation:

$$\theta_c = \sin^{-1}(n_2/n_1).$$

Substituting the identified values gives

Equation:

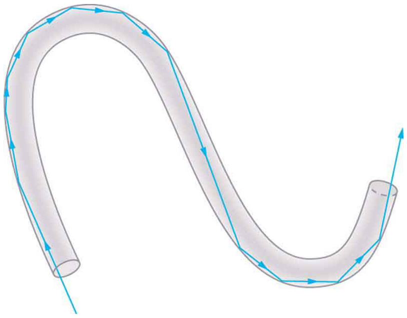
$$\theta_c = \sin^{-1}(1.00/1.49) = \sin^{-1}(0.671)$$
$$42.2^{\text{o}}.$$

Discussion

This means that any ray of light inside the plastic that strikes the surface at an angle greater than 42.2^{o} will be totally reflected. This will make the inside surface of the clear plastic a perfect mirror for such rays without any need for the silvering used on common mirrors. Different combinations of materials have different critical angles, but any combination with $n_1 > n_2$ can produce total internal reflection. The same calculation as made here shows that the critical angle for a ray going from water to air is 48.6^{o}, while that from diamond to air is 24.4^{o}, and that from flint glass to crown glass is 66.3^{o}. There is no total reflection for rays going in the other direction—for example, from air to water—since the condition that the second medium must have a smaller index of refraction is not satisfied. A number of interesting applications of total internal reflection follow.

Fiber Optics: Endoscopes to Telephones

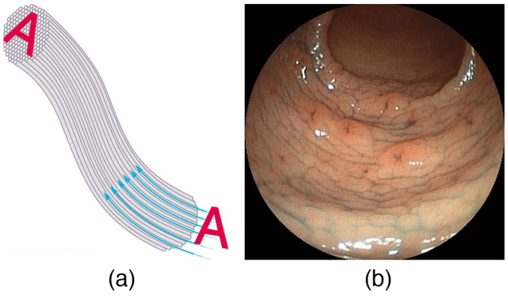
Fiber optics is one application of total internal reflection that is in wide use. In communications, it is used to transmit telephone, internet, and cable TV signals. **Fiber optics** employs the transmission of light down fibers of plastic or glass. Because the fibers are thin, light entering one is likely to strike the inside surface at an angle greater than the critical angle and, thus, be totally reflected (See [link].) The index of refraction outside the fiber must be smaller than inside, a condition that is easily satisfied by coating the outside of the fiber with a material having an appropriate refractive index. In fact, most fibers have a varying refractive index to allow more light to be guided along the fiber through total internal refraction. Rays are reflected around corners as shown, making the fibers into tiny light pipes.



Light entering a thin fiber may strike the inside surface at large or grazing angles and is completely reflected if these angles exceed the critical angle. Such rays continue down the fiber, even following it around corners, since the angles of reflection and incidence remain large.

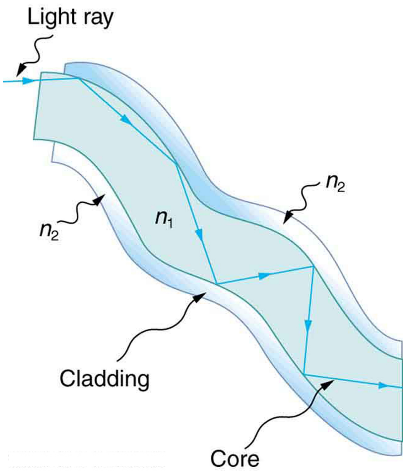
Bundles of fibers can be used to transmit an image without a lens, as illustrated in [link]. The output of a device called an **endoscope** is shown in [link](b). Endoscopes are used to explore the body through various orifices or minor incisions. Light is transmitted down one fiber bundle to illuminate internal parts, and the reflected light is transmitted back out through another to be observed. Surgery can be performed, such as arthroscopic surgery on the knee joint, employing cutting tools attached to and observed with the endoscope. Samples can also be obtained, such as by lassoing an intestinal polyp for external examination.

Fiber optics has revolutionized surgical techniques and observations within the body. There are a host of medical diagnostic and therapeutic uses. The flexibility of the fiber optic bundle allows it to navigate around difficult and small regions in the body, such as the intestines, the heart, blood vessels, and joints. Transmission of an intense laser beam to burn away obstructing plaques in major arteries as well as delivering light to activate chemotherapy drugs are becoming commonplace. Optical fibers have in fact enabled microsurgery and remote surgery where the incisions are small and the surgeon’s fingers do not need to touch the diseased tissue.



(a) An image is transmitted by a bundle of fibers that have fixed neighbors. (b) An endoscope is used to probe the body, both transmitting light to the interior and returning an image such as the one shown. (credit: Med_Chaos, Wikimedia Commons)

Fibers in bundles are surrounded by a cladding material that has a lower index of refraction than the core. (See [link].) The cladding prevents light from being transmitted between fibers in a bundle. Without cladding, light could pass between fibers in contact, since their indices of refraction are identical. Since no light gets into the cladding (there is total internal reflection back into the core), none can be transmitted between clad fibers that are in contact with one another. The cladding prevents light from escaping out of the fiber; instead most of the light is propagated along the length of the fiber, minimizing the loss of signal and ensuring that a quality image is formed at the other end. The cladding and an additional protective layer make optical fibers flexible and durable.



Fibers in bundles are clad by a material that has a lower index of refraction than the core to ensure total internal reflection, even when fibers are in contact with one another. This shows a single fiber with its cladding.

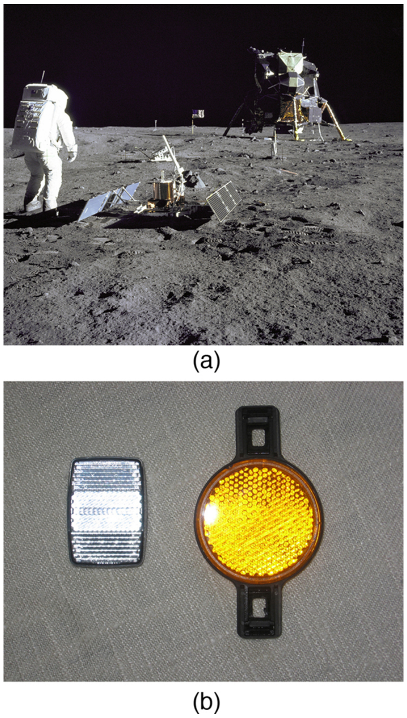
Note:
Cladding
The cladding prevents light from being transmitted between fibers in a bundle.

Special tiny lenses that can be attached to the ends of bundles of fibers are being designed and fabricated. Light emerging from a fiber bundle can be focused and a tiny spot can be imaged. In some cases the spot can be scanned, allowing quality imaging of a region inside the body. Special minute optical filters inserted at the end of the fiber bundle have the capacity to image tens of microns below the surface without cutting the surface—non-intrusive diagnostics. This is particularly useful for determining the extent of cancers in the stomach and bowel.

Most telephone conversations and Internet communications are now carried by laser signals along optical fibers. Extensive optical fiber cables have been placed on the ocean floor and underground to enable optical communications. Optical fiber communication systems offer several advantages over electrical (copper) based systems, particularly for long distances. The fibers can be made so transparent that light can travel many kilometers before it becomes dim enough to require amplification—much superior to copper conductors. This property of optical fibers is called *low loss*. Lasers emit light with characteristics that allow far more conversations in one fiber than are possible with electric signals on a single conductor. This property of optical fibers is called *high bandwidth*. Optical signals in one fiber do not produce undesirable effects in other adjacent fibers. This property of optical fibers is called *reduced crosstalk*. We shall explore the unique characteristics of laser radiation in a later chapter.

Corner Reflectors and Diamonds

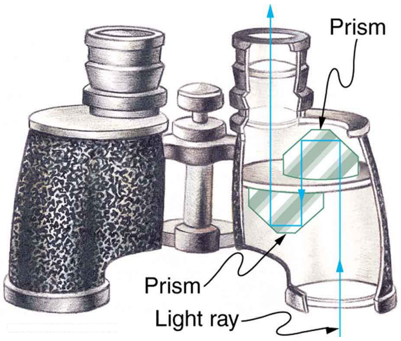
A light ray that strikes an object consisting of two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came. This is true whenever the reflecting surfaces are perpendicular, and it is independent of the angle of incidence. Such an object, shown in [link], is called a **corner reflector**, since the light bounces from its inside corner. Many inexpensive reflector buttons on bicycles, cars, and warning signs have corner reflectors designed to return light in the direction from which it originated. It was more expensive for astronauts to place one on the moon. Laser signals can be bounced from that corner reflector to measure the gradually increasing distance to the moon with great precision.


(a)

(b)

(a) Astronauts placed a corner reflector on the moon to measure its gradually increasing orbital distance. (credit: NASA) (b) The bright spots on these bicycle safety reflectors are reflections of the flash of the camera that took this picture on a dark night. (credit: Julo, Wikimedia Commons)

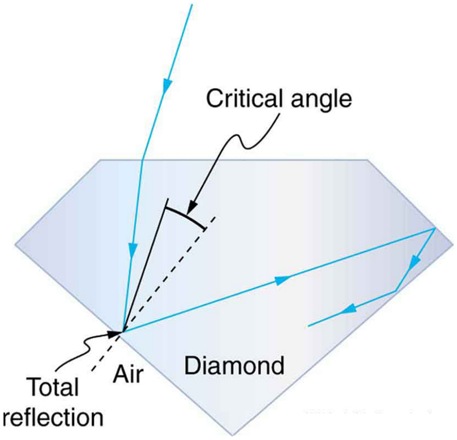
Corner reflectors are perfectly efficient when the conditions for total internal reflection are satisfied. With common materials, it is easy to obtain a critical angle that is less than 45°. One use of these perfect mirrors is in binoculars, as shown in [link]. Another use is in periscopes found in submarines.



These binoculars employ corner reflectors with total internal reflection to get light to the observer's eyes.

The Sparkle of Diamonds

Total internal reflection, coupled with a large index of refraction, explains why diamonds sparkle more than other materials. The critical angle for a diamond-to-air surface is only 24.4°, and so when light enters a diamond, it has trouble getting back out. (See [link].) Although light freely enters the diamond, it can exit only if it makes an angle less than 24.4°. Facets on diamonds are specifically intended to make this unlikely, so that the light can exit only in certain places. Good diamonds are very clear, so that the light makes many internal reflections and is concentrated at the few places it can exit—hence the sparkle. (Zircon is a natural gemstone that has an exceptionally large index of refraction, but not as large as diamond, so it is not as highly prized. Cubic zirconia is manufactured and has an even higher index of refraction (≈ 2.17), but still less than that of diamond.) The colors you see emerging from a sparkling diamond are not due to the diamond's color, which is usually nearly colorless. Those colors result from dispersion, the topic of Dispersion: The Rainbow and Prisms. Colored diamonds get their color from structural defects of the crystal lattice and the inclusion of minute quantities of graphite and other materials. The Argyle Mine in Western Australia produces around 90% of the world's pink, red, champagne, and cognac diamonds, while around 50% of the world's clear diamonds come from central and southern Africa.



Light cannot easily escape a diamond, because its critical angle with air is so small. Most reflections are total, and the facets are placed so that light can exit only in particular ways—thus concentrating the light and making the diamond sparkle.

Note:
PhET Explorations: Bending Light
Explore bending of light between two media with different indices of refraction. See how changing from air to water to glass changes the bending angle. Play with prisms of different shapes and make rainbows.

https://phet.colorado.edu/sims/html/bending-light/latest/bending-light_en.html

Section Summary

- The incident angle that produces an angle of refraction of 90° is called critical angle.
- Total internal reflection is a phenomenon that occurs at the boundary between two mediums, such that if the incident angle in the first medium is greater than the critical angle, then all the light is reflected back into that medium.
- Fiber optics involves the transmission of light down fibers of plastic or glass, applying the principle of total internal reflection.
- Endoscopes are used to explore the body through various orifices or minor incisions, based on the transmission of light through optical fibers.
- Cladding prevents light from being transmitted between fibers in a bundle.
- Diamonds sparkle due to total internal reflection coupled with a large index of refraction.

Conceptual Questions

Exercise:

Problem:

A ring with a colorless gemstone is dropped into water. The gemstone becomes invisible when submerged. Can it be a diamond? Explain.

Exercise:

Problem:

A high-quality diamond may be quite clear and colorless, transmitting all visible wavelengths with little absorption. Explain how it can sparkle with flashes of brilliant color when illuminated by white light.

Exercise:

Problem:

Is it possible that total internal reflection plays a role in rainbows? Explain in terms of indices of refraction and angles, perhaps referring to [link]. Some of us have seen the formation of a double rainbow. Is it physically possible to observe a triple rainbow?

Double rainbows are not a very common observance. (credit: InvictusOU812, Flickr)

Exercise:

Problem:

The most common type of mirage is an illusion that light from faraway objects is reflected by a pool of water that is not really there. Mirages are generally observed in deserts, when there is a hot layer of air near the ground. Given that the refractive index of air is lower for air at higher temperatures, explain how mirages can be formed.

Problems & Exercises

Exercise:

Problem:

Verify that the critical angle for light going from water to air is 48.6°, as discussed at the end of [link], regarding the critical angle for light traveling in a polystyrene (a type of plastic) pipe surrounded by air.

Exercise:

Problem:

(a) At the end of [link], it was stated that the critical angle for light going from diamond to air is 24.4°. Verify this. (b) What is the critical angle for light going from zircon to air?

Exercise:

Problem:

An optical fiber uses flint glass clad with crown glass. What is the critical angle?

Solution:

66.3°

Exercise:

Problem:

At what minimum angle will you get total internal reflection of light traveling in water and reflected from ice?

Exercise:

Problem:

Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0^{o}, what must be the minimum index of refraction of the material from which the reflector is made?

Solution:

> 1.414

Exercise:

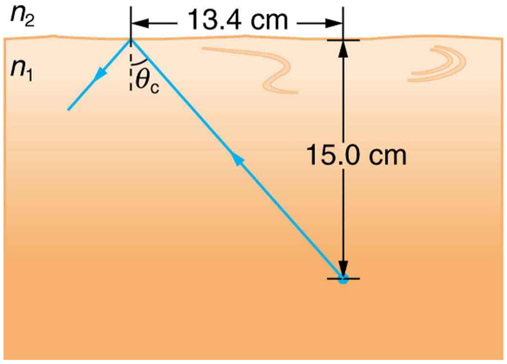
Problem:

You can determine the index of refraction of a substance by determining its critical angle. (a) What is the index of refraction of a substance that has a critical angle of 68.4^{o} when submerged in water? What is the substance, based on [link]? (b) What would the critical angle be for this substance in air?

Exercise:

Problem:

A ray of light, emitted beneath the surface of an unknown liquid with air above it, undergoes total internal reflection as shown in [link]. What is the index of refraction for the liquid and its likely identification?



A light ray inside a liquid strikes the surface at the critical angle and undergoes total internal reflection.

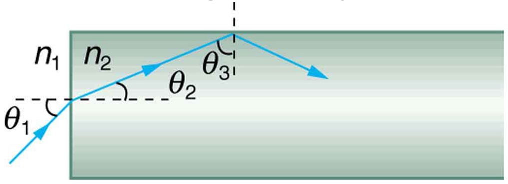
Solution:

1.50, benzene

Exercise:

Problem:

A light ray entering an optical fiber surrounded by air is first refracted and then reflected as shown in [link]. Show that if the fiber is made from crown glass, any incident ray will be totally internally reflected.



A light ray enters the end of a fiber, the surface of which is perpendicular to its sides. Examine the conditions under which it

may be totally internally
reflected.

Glossary

critical angle
: incident angle that produces an angle of refraction of 90º

fiber optics
: transmission of light down fibers of plastic or glass, applying the principle of total internal reflection

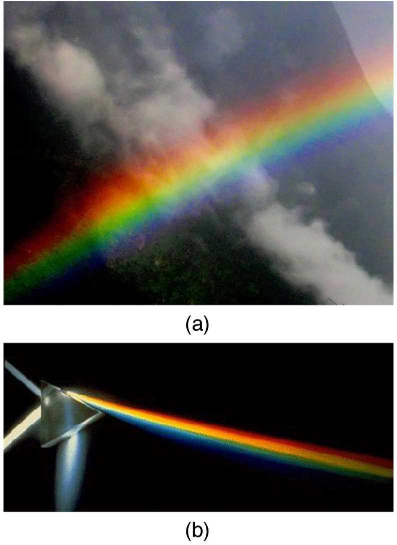
corner reflector
: an object consisting of two mutually perpendicular reflecting surfaces, so that the light that enters is reflected back exactly parallel to the direction from which it came

zircon
: natural gemstone with a large index of refraction

Dispersion: The Rainbow and Prisms

- Explain the phenomenon of dispersion and discuss its advantages and disadvantages.

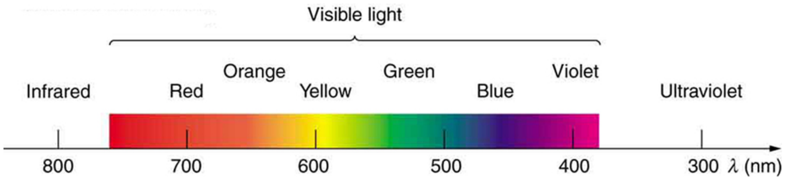
Everyone enjoys the spectacle of a rainbow glimmering against a dark stormy sky. How does sunlight falling on clear drops of rain get broken into the rainbow of colors we see? The same process causes white light to be broken into colors by a clear glass prism or a diamond. (See [link].)



The colors of the rainbow (a) and those produced by a prism (b) are identical. (credit: Alfredo55, Wikimedia Commons; NASA)

We see about six colors in a rainbow—red, orange, yellow, green, blue, and violet; sometimes indigo is listed, too. Those colors are associated with different wavelengths of light, as shown in [link]. When our eye receives pure-wavelength light, we tend to see only one of the six colors, depending on wavelength. The thousands of other hues we can sense in other situations are our eye’s response to various mixtures of wavelengths. White light, in particular, is a fairly uniform mixture of all visible wavelengths. Sunlight, considered to be white, actually appears to be a bit yellow because of its mixture of wavelengths, but it does contain all visible wavelengths. The sequence of colors in rainbows is the same sequence as the colors plotted versus wavelength in [link]. What this implies is that white light is spread out according to wavelength in a rainbow. **Dispersion** is defined as the spreading of white light into its full spectrum of wavelengths. More technically, dispersion occurs whenever there is a process that changes the direction of light in a manner that depends on wavelength. Dispersion, as a general phenomenon, can occur for any type of wave and always involves wavelength-dependent processes.

Note:
Dispersion
Dispersion is defined to be the spreading of white light into its full spectrum of wavelengths.



Even though rainbows are associated with seven colors, the rainbow is a continuous distribution of colors according to wavelengths.

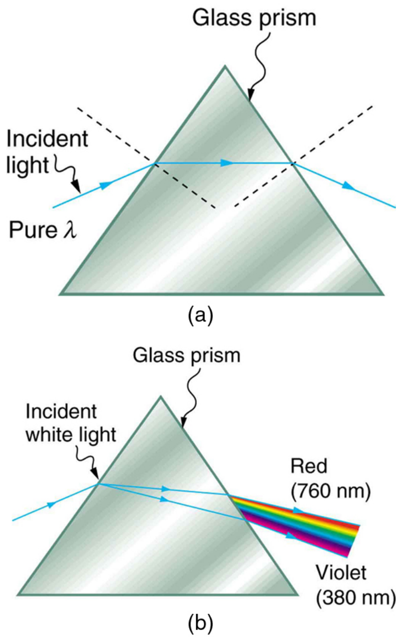
Refraction is responsible for dispersion in rainbows and many other situations. The angle of refraction depends on the index of refraction, as we saw in The Law of Refraction. We know that the index of refraction n depends on the medium. But for a given medium, n also depends on wavelength. (See [link]. Note that, for a given medium, n increases as wavelength decreases and is greatest for violet light. Thus violet light is bent more than red light, as shown for a prism in [link](b), and the light is dispersed into the same sequence of wavelengths as seen in [link] and [link].

Note:
Making Connections: Dispersion
Any type of wave can exhibit dispersion. Sound waves, all types of electromagnetic waves, and water waves can be dispersed according to wavelength. Dispersion occurs whenever the speed of propagation depends on wavelength, thus separating and spreading out various wavelengths. Dispersion may require special circumstances and can result in spectacular displays such as in the production of a rainbow. This is also

true for sound, since all frequencies ordinarily travel at the same speed. If you listen to sound through a long tube, such as a vacuum cleaner hose, you can easily hear it is dispersed by interaction with the tube. Dispersion, in fact, can reveal a great deal about what the wave has encountered that disperses its wavelengths. The dispersion of electromagnetic radiation from outer space, for example, has revealed much about what exists between the stars—the so-called empty space.

Medium	**Red (660 nm)**	**Orange (610 nm)**	**Yellow (580 nm)**	**Green (550 nm)**	**Blue (470 nm)**	**Violet (410 nm)**
Water	1.331	1.332	1.333	1.335	1.338	1.342
Diamond	2.410	2.415	2.417	2.426	2.444	2.458
Glass, crown	1.512	1.514	1.518	1.519	1.524	1.530
Glass, flint	1.662	1.665	1.667	1.674	1.684	1.698
Polystyrene	1.488	1.490	1.492	1.493	1.499	1.506
Quartz, fused	1.455	1.456	1.458	1.459	1.462	1.468

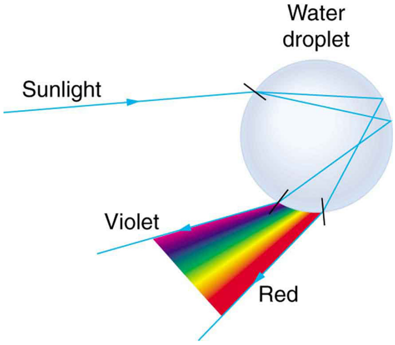
Index of Refraction *n* in Selected Media at Various Wavelengths



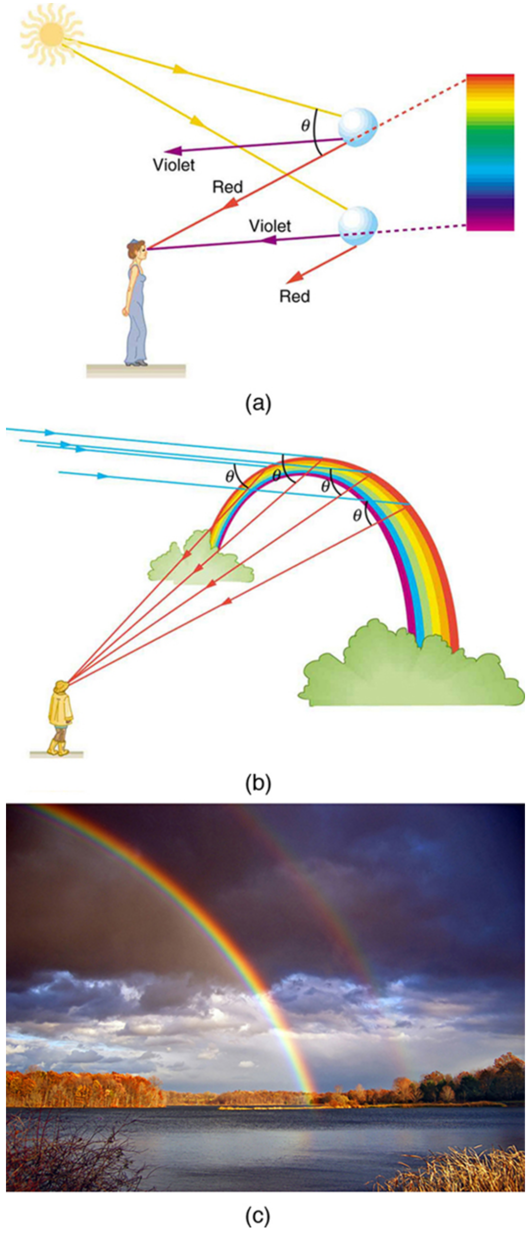
(a) A pure wavelength of light falls onto a prism and is refracted at both surfaces. (b) White light is dispersed by the prism (shown exaggerated). Since the index of refraction varies with wavelength, the angles of refraction vary with wavelength. A sequence of red to violet is produced, because the index of refraction increases steadily with decreasing wavelength.

Rainbows are produced by a combination of refraction and reflection. You may have noticed that you see a rainbow only when you look away from the sun. Light enters a drop of water and is reflected from the back of the drop, as shown in [link]. The light is refracted both as it enters and as it leaves the drop. Since the index of refraction of water varies with wavelength, the light is dispersed, and a rainbow is observed, as shown in [link] (a). (There is no dispersion caused by reflection at the back surface, since the law of reflection does not depend on wavelength.) The actual rainbow of colors seen by an observer depends on the myriad of rays being refracted and reflected toward the observer's eyes from numerous drops of water. The effect is most spectacular when the background is dark, as in stormy weather, but can also be observed in waterfalls and lawn sprinklers. The arc of a rainbow comes from the need to be looking at a specific angle relative to the direction of the sun, as illustrated in [link] (b). (If there are two reflections of light within the water drop, another "secondary" rainbow is produced. This rare event produces an arc that lies above the primary rainbow arc—see [link] (c).)

Note:
Rainbows
Rainbows are produced by a combination of refraction and reflection.



Part of the light falling on this water drop enters and is reflected from the back of the drop. This light is refracted and dispersed both as it enters and as it leaves the drop.



(a)

(b)

(c)

(a) Different colors emerge in different directions, and so you must look at different locations to see the various colors of a rainbow. (b) The arc of a rainbow results from the fact that a line between the observer and any point on the arc must make the correct angle with the parallel rays of sunlight to receive the refracted rays. (c)

Double rainbow. (credit: Nicholas, Wikimedia Commons)

Dispersion may produce beautiful rainbows, but it can cause problems in optical systems. White light used to transmit messages in a fiber is dispersed, spreading out in time and eventually overlapping with other messages. Since a laser produces a nearly pure wavelength, its light experiences little dispersion, an advantage over white light for transmission of information. In contrast, dispersion of electromagnetic waves coming to us from outer space can be used to determine the amount of matter they pass through. As with many phenomena, dispersion can be useful or a nuisance, depending on the situation and our human goals.

Note:
PhET Explorations: Geometric Optics
How does a lens form an image? See how light rays are refracted by a lens. Watch how the image changes when you adjust the focal length of the lens, move the object, move the lens, or move the screen.

https://phet.colorado.edu/sims/geometric-optics/geometric-optics_en.html

Section Summary

- The spreading of white light into its full spectrum of wavelengths is called dispersion.
- Rainbows are produced by a combination of refraction and reflection and involve the dispersion of sunlight into a continuous distribution of colors.
- Dispersion produces beautiful rainbows but also causes problems in certain optical systems.

Problems & Exercises

Exercise:

Problem:

(a) What is the ratio of the speed of red light to violet light in diamond, based on [link]? (b) What is this ratio in polystyrene? (c) Which is more dispersive?

Exercise:

Problem:

A beam of white light goes from air into water at an incident angle of 75.0°. At what angles are the red (660 nm) and violet (410 nm) parts of the light refracted?

Solution:

46.5°, red; 46.0°, violet

Exercise:

Problem:

By how much do the critical angles for red (660 nm) and violet (410 nm) light differ in a diamond surrounded by air?

Exercise:

Problem:

(a) A narrow beam of light containing yellow (580 nm) and green (550 nm) wavelengths goes from polystyrene to air, striking the surface at a 30.0° incident angle. What is the angle between the colors when they emerge? (b) How far would they have to travel to be separated by 1.00 mm?

Solution:

(a) 0.043°

(b) 1.33 m

Exercise:

Problem:

A parallel beam of light containing orange (610 nm) and violet (410 nm) wavelengths goes from fused quartz to water, striking the surface between them at a 60.0° incident angle. What is the angle between the two colors in water?

Exercise:

Problem:

A ray of 610 nm light goes from air into fused quartz at an incident angle of 55.0°. At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?

Solution:

71.3º

Exercise:

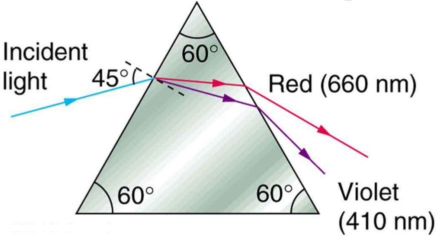
Problem:

A narrow beam of light containing red (660 nm) and blue (470 nm) wavelengths travels from air through a 1.00 cm thick flat piece of crown glass and back to air again. The beam strikes at a 30.0º incident angle. (a) At what angles do the two colors emerge? (b) By what distance are the red and blue separated when they emerge?

Exercise:

Problem:

A narrow beam of white light enters a prism made of crown glass at a 45.0º incident angle, as shown in [link]. At what angles, θ_R and θ_V, do the red (660 nm) and violet (410 nm) components of the light emerge from the prism?



This prism will disperse the white light into a rainbow of colors. The incident angle is 45.0º, and the angles at which the red and violet light emerge are θ_R and θ_V.

Solution:

53.5º, red; 55.2º, violet

Glossary

dispersion

spreading of white light into its full spectrum of wavelengths

rainbow

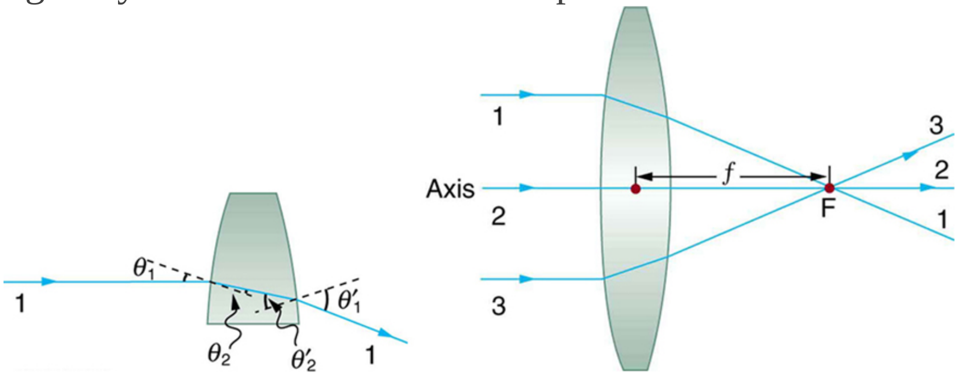
dispersion of sunlight into a continuous distribution of colors according to wavelength, produced by the refraction and reflection of sunlight by water droplets in the sky

Image Formation by Lenses

- List the rules for ray tracking for thin lenses.
- Illustrate the formation of images using the technique of ray tracking.
- Determine power of a lens given the focal length.

Lenses are found in a huge array of optical instruments, ranging from a simple magnifying glass to the eye to a camera’s zoom lens. In this section, we will use the law of refraction to explore the properties of lenses and how they form images.

The word *lens* derives from the Latin word for a lentil bean, the shape of which is similar to the convex lens in [link]. The convex lens shown has been shaped so that all light rays that enter it parallel to its axis cross one another at a single point on the opposite side of the lens. (The axis is defined to be a line normal to the lens at its center, as shown in [link].) Such a lens is called a **converging (or convex) lens** for the converging effect it has on light rays. An expanded view of the path of one ray through the lens is shown, to illustrate how the ray changes direction both as it enters and as it leaves the lens. Since the index of refraction of the lens is greater than that of air, the ray moves towards the perpendicular as it enters and away from the perpendicular as it leaves. (This is in accordance with the law of refraction.) Due to the lens’s shape, light is thus bent toward the axis at both surfaces. The point at which the rays cross is defined to be the **focal point** F of the lens. The distance from the center of the lens to its focal point is defined to be the **focal length** f of the lens. [link] shows how a converging lens, such as that in a magnifying glass, can converge the nearly parallel light rays from the sun to a small spot.



Rays of light entering a converging lens parallel to its axis converge at its focal point F. (Ray 2 lies on the axis of the lens.) The distance from the center of the lens to the focal point is the lens's focal length f. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Note:
Converging or Convex Lens
The lens in which light rays that enter it parallel to its axis cross one another at a single point on the opposite side with a converging effect is called converging lens.

Note:
Focal Point F
The point at which the light rays cross is called the focal point F of the lens.

Note:
Focal Length f
The distance from the center of the lens to its focal point is called focal length f.

Sunlight focused by a converging magnifying glass can burn paper. Light rays from the sun are nearly parallel and cross at the focal point of the lens. The more powerful the lens, the closer to the lens the rays will cross.

The greater effect a lens has on light rays, the more powerful it is said to be. For example, a powerful converging lens will focus parallel light rays closer to itself and will have a smaller focal length than a weak lens. The light will also focus into a smaller and more intense spot for a more powerful lens. The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

Equation:

$$P=\frac{1}{f}.$$

Note:

Power P

The **power** P of a lens is defined to be the inverse of its focal length. In equation form, this is

Equation:

$$P = \frac{1}{f}.$$

where f is the focal length of the lens, which must be given in meters (and not cm or mm). The power of a lens P has the unit diopters (D), provided that the focal length is given in meters. That is, $1\ \text{D} = 1/\text{m}$, or $1\ \text{m}^{-1}$. (Note that this power (optical power, actually) is not the same as power in watts defined in Work, Energy, and Energy Resources. It is a concept related to the effect of optical devices on light.) Optometrists prescribe common spectacles and contact lenses in units of diopters.

Example:

What is the Power of a Common Magnifying Glass?

Suppose you take a magnifying glass out on a sunny day and you find that it concentrates sunlight to a small spot 8.00 cm away from the lens. What are the focal length and power of the lens?

Strategy

The situation here is the same as those shown in [link] and [link]. The Sun is so far away that the Sun’s rays are nearly parallel when they reach Earth. The magnifying glass is a convex (or converging) lens, focusing the nearly parallel rays of sunlight. Thus the focal length of the lens is the distance from the lens to the spot, and its power is the inverse of this distance (in m).

Solution

The focal length of the lens is the distance from the center of the lens to the spot, given to be 8.00 cm. Thus,

Equation:

$$f = 8.00\ \text{cm}.$$

To find the power of the lens, we must first convert the focal length to meters; then, we substitute this value into the equation for power. This gives

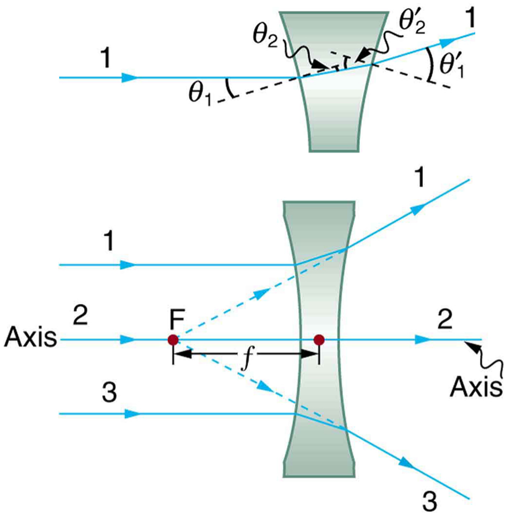
Equation:

$$P=\frac{1}{f}=\frac{1}{0.0800\text{ m}}=12.5\text{ D}.$$

Discussion

This is a relatively powerful lens. The power of a lens in diopters should not be confused with the familiar concept of power in watts. It is an unfortunate fact that the word “power” is used for two completely different concepts. If you examine a prescription for eyeglasses, you will note lens powers given in diopters. If you examine the label on a motor, you will note energy consumption rate given as a power in watts.

[link] shows a concave lens and the effect it has on rays of light that enter it parallel to its axis (the path taken by ray 2 in the figure is the axis of the lens). The concave lens is a **diverging lens**, because it causes the light rays to bend away (diverge) from its axis. In this case, the lens has been shaped so that all light rays entering it parallel to its axis appear to originate from the same point, F, defined to be the focal point of a diverging lens. The distance from the center of the lens to the focal point is again called the focal length f of the lens. Note that the focal length and power of a diverging lens are defined to be negative. For example, if the distance to F in [link] is 5.00 cm, then the focal length is $f=\text{–5.00 cm}$ and the power of the lens is $P=\text{–20 D}$. An expanded view of the path of one ray through the lens is shown in the figure to illustrate how the shape of the lens, together with the law of refraction, causes the ray to follow its particular path and be diverged.

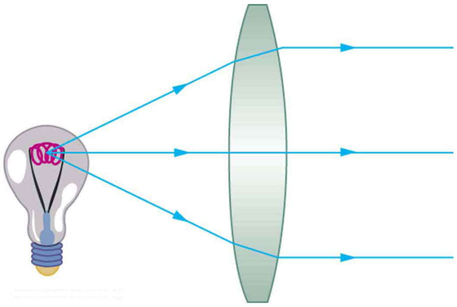


Rays of light entering a diverging lens parallel to its axis are diverged, and all appear to originate at its focal point F. The dashed lines are not rays—they indicate the directions from which the rays appear to come. The focal length f of a diverging lens is negative. An expanded view of the path taken by ray 1 shows the perpendiculars and the angles of incidence and refraction at both surfaces.

Note:
Diverging Lens

A lens that causes the light rays to bend away from its axis is called a diverging lens.

As noted in the initial discussion of the law of refraction in The Law of Refraction, the paths of light rays are exactly reversible. This means that the direction of the arrows could be reversed for all of the rays in [link] and [link]. For example, if a point light source is placed at the focal point of a convex lens, as shown in [link], parallel light rays emerge from the other side.



A small light source, like a light bulb filament, placed at the focal point of a convex lens, results in parallel rays of light emerging from the other side. The paths are exactly the reverse of those shown in [link]. This technique is used in lighthouses and sometimes in traffic lights to produce a directional beam of light from a source that emits light in all directions.

Ray Tracing and Thin Lenses

Ray tracing is the technique of determining or following (tracing) the paths that light rays take. For rays passing through matter, the law of refraction is used to trace the paths. Here we use ray tracing to help us understand the action of lenses in situations ranging from forming images on film to magnifying small print to correcting nearsightedness. While ray tracing for complicated lenses, such as those found in sophisticated cameras, may require computer techniques, there is a set of simple rules for tracing rays through thin lenses. A **thin lens** is defined to be one whose thickness allows rays to refract, as illustrated in [link], but does not allow properties such as dispersion and aberrations. An ideal thin lens has two refracting surfaces but the lens is thin enough to assume that light rays bend only once. A thin symmetrical lens has two focal points, one on either side and both at the same distance from the lens. (See [link].) Another important characteristic of a thin lens is that light rays through its center are deflected by a negligible amount, as seen in [link].

Note:

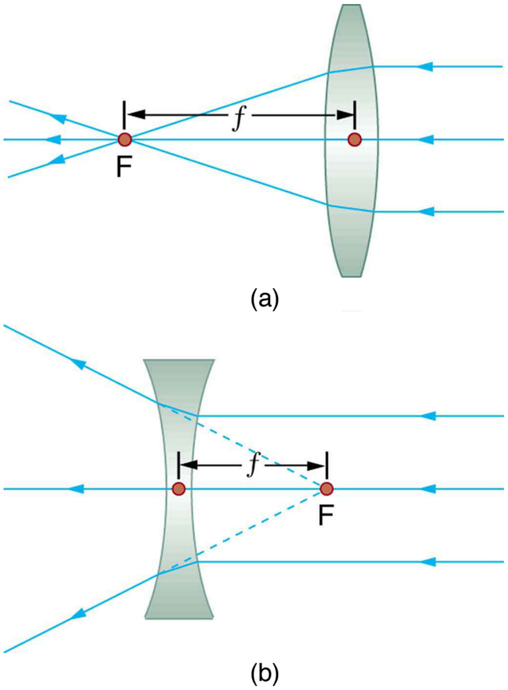
Thin Lens

A thin lens is defined to be one whose thickness allows rays to refract but does not allow properties such as dispersion and aberrations.

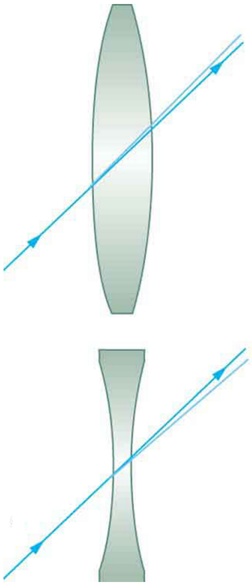
Note:

Take-Home Experiment: A Visit to the Optician

Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they act like thin lenses.



Thin lenses have the same focal length on either side. (a) Parallel light rays entering a converging lens from the right cross at its focal point on the left. (b) Parallel light rays entering a diverging lens from the right seem to come from the focal point on the right.



The light ray through the center of a thin lens is deflected by a negligible amount and is assumed to emerge parallel to its original path (shown as a shaded line).

Using paper, pencil, and a straight edge, ray tracing can accurately describe the operation of a lens. The rules for ray tracing for thin lenses are based on the illustrations already discussed:

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side. (See rays 1 and 3 in [link].)
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F. (See rays 1 and 3 in [link].)
3. A ray passing through the center of either a converging or a diverging lens does not change direction. (See [link], and see ray 2 in [link] and [link].)
4. A ray entering a converging lens through its focal point exits parallel to its axis. (The reverse of rays 1 and 3 in [link].)
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis. (The reverse of rays 1 and 3 in [link].)

Note:

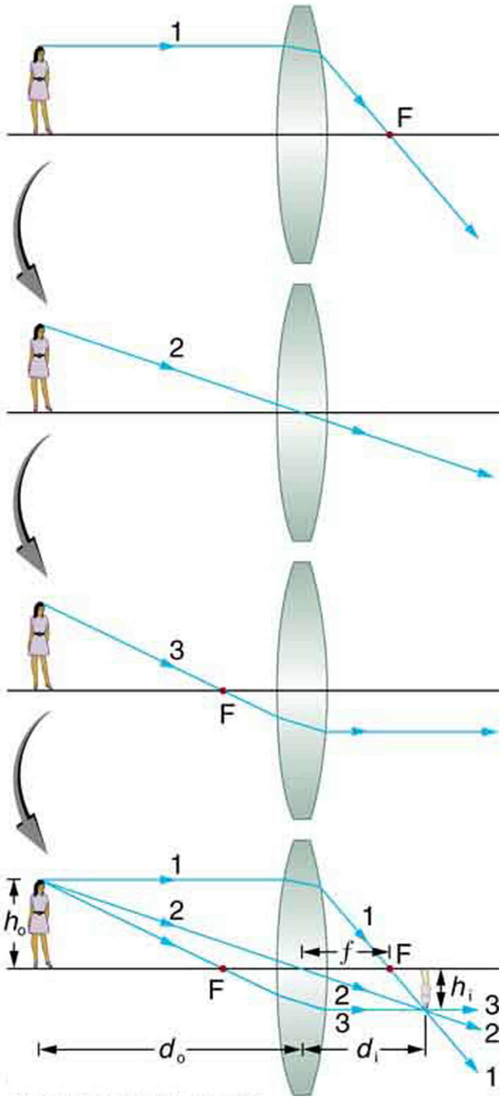
Rules for Ray Tracing

1. A ray entering a converging lens parallel to its axis passes through the focal point F of the lens on the other side.
2. A ray entering a diverging lens parallel to its axis seems to come from the focal point F.
3. A ray passing through the center of either a converging or a diverging lens does not change direction.
4. A ray entering a converging lens through its focal point exits parallel to its axis.
5. A ray that enters a diverging lens by heading toward the focal point on the opposite side exits parallel to the axis.

Image Formation by Thin Lenses

In some circumstances, a lens forms an obvious image, such as when a movie projector casts an image onto a screen. In other cases, the image is less obvious. Where, for example, is the image formed by eyeglasses? We will use ray tracing for thin lenses to illustrate how they form images, and we will develop equations to describe the image formation quantitatively.

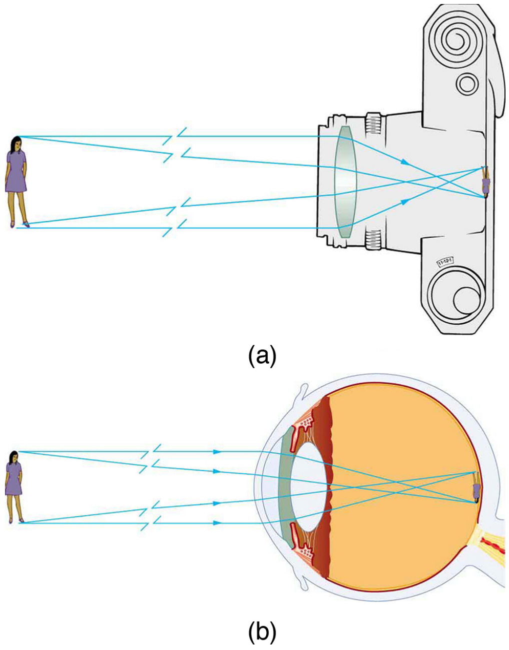
Consider an object some distance away from a converging lens, as shown in [link]. To find the location and size of the image formed, we trace the paths of selected light rays originating from one point on the object, in this case the top of the person’s head. The figure shows three rays from the top of the object that can be traced using the ray tracing rules given above. (Rays leave this point going in many directions, but we concentrate on only a few with paths that are easy to trace.) The first ray is one that enters the lens parallel to its axis and passes through the focal point on the other side (rule 1). The second ray passes through the center of the lens without changing direction (rule 3). The third ray passes through the nearer focal point on its way into the lens and leaves the lens parallel to its axis (rule 4). The three rays cross at the same point on the other side of the lens. The image of the top of the person’s head is located at this point. All rays that come from the same point on the top of the person’s head are refracted in such a way as to cross at the point shown. Rays from another point on the object, such as her belt buckle, will also cross at another common point, forming a complete image, as shown. Although three rays are traced in [link], only two are necessary to locate the image. It is best to trace rays for which there are simple ray tracing rules. Before applying ray tracing to other situations, let us consider the example shown in [link] in more detail.



Ray tracing is used to locate the image formed by a lens. Rays originating from the same point on the object are traced—the three chosen rays each follow one of the rules for ray tracing, so that their paths are easy to determine. The image is located at the point where the rays cross. In this case, a real image—one that can be projected on a screen—is formed.

The image formed in [link] is a **real image**, meaning that it can be projected. That is, light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye, for example. [link] shows how such an image would be projected onto film by a camera lens. This figure also shows how a real image is projected onto the retina by the lens of an eye. Note that the image is there whether it is projected onto a screen or not.

Note:
Real Image
The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.



Real images can be projected. (a) A real image of the person is projected onto film. (b) The converging nature of the multiple surfaces that make up the eye result in the projection of a real image on the retina.

Several important distances appear in [link]. We define d_o to be the object distance, the distance of an object from the center of a lens. **Image distance** d_i is defined to be the distance of the image from the center of a lens. The height of the object and height of the image are given the symbols h_o and h_i , respectively. Images that appear upright relative to the object have heights that are positive and those that are inverted have negative heights. Using the rules of ray tracing and making a scale drawing with paper and pencil, like that in [link], we can accurately describe the location and size of an image. But the real benefit of ray tracing is in visualizing how images are formed in a variety of situations. To obtain numerical information, we use a pair of equations that can be derived from a geometric analysis of ray tracing for thin lenses. The **thin lens equations** are

Equation:

$$\frac{1}{d_{\text{o}}}+\frac{1}{d_{\text{i}}}=\frac{1}{f}$$

and

Equation:

$$\frac{h_{\text{i}}}{h_{\text{o}}}=-\frac{d_{\text{i}}}{d_{\text{o}}}=m.$$

We define the ratio of image height to object height ($h_{\text{i}}/h_{\text{o}}$) to be the **magnification** m. (The minus sign in the equation above will be discussed shortly.) The thin lens equations are broadly applicable to all situations involving thin lenses (and “thin” mirrors, as we will see later). We will explore many features of image formation in the following worked examples.

Note:

Image Distance

The distance of the image from the center of the lens is called image distance.

Note:

Thin Lens Equations and Magnification

Equation:

$$\frac{1}{d_{\text{o}}}+\frac{1}{d_{\text{i}}}=\frac{1}{f}$$

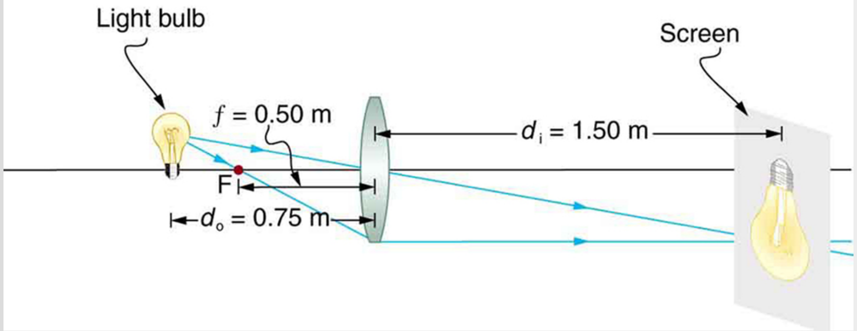
Equation:

$$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = m$$

Example:

Finding the Image of a Light Bulb Filament by Ray Tracing and by the Thin Lens Equations

A clear glass light bulb is placed 0.750 m from a convex lens having a 0.500 m focal length, as shown in [link]. Use ray tracing to get an approximate location for the image. Then use the thin lens equations to calculate (a) the location of the image and (b) its magnification. Verify that ray tracing and the thin lens equations produce consistent results.



A light bulb placed 0.750 m from a lens having a 0.500 m focal length produces a real image on a poster board as discussed in the example above. Ray tracing predicts the image location and size.

Strategy and Concept

Since the object is placed farther away from a converging lens than the focal length of the lens, this situation is analogous to those illustrated in [link] and [link]. Ray tracing to scale should produce similar results for d_i. Numerical solutions for d_i and m can be obtained using the thin lens equations, noting that $d_\text{o} = 0.750\text{ m}$ and $f = 0.500\text{ m}$.

Solutions (Ray tracing)

The ray tracing to scale in [link] shows two rays from a point on the bulb's filament crossing about 1.50 m on the far side of the lens. Thus the image distance d_{i} is about 1.50 m. Similarly, the image height based on ray tracing is greater than the object height by about a factor of 2, and the image is inverted. Thus m is about –2. The minus sign indicates that the image is inverted.

The thin lens equations can be used to find d_{i} from the given information:

Equation:

$$\frac{1}{d_{\text{o}}}+\frac{1}{d_{\text{i}}}=\frac{1}{f}.$$

Rearranging to isolate d_{i} gives

Equation:

$$\frac{1}{d_{\text{i}}}=\frac{1}{f}-\frac{1}{d_{\text{o}}}.$$

Entering known quantities gives a value for $1/d_{\text{i}}$:

Equation:

$$\frac{1}{d_{\text{i}}}=\frac{1}{0.500\text{ m}}-\frac{1}{0.750\text{ m}}=\frac{0.667}{\text{m}}.$$

This must be inverted to find d_{i}:

Equation:

$$d_{\text{i}}=\frac{\text{m}}{0.667}=1.50\text{ m}.$$

Note that another way to find d_{i} is to rearrange the equation:

Equation:

$$\frac{1}{d_{\text{i}}}=\frac{1}{f}-\frac{1}{d_{\text{o}}}.$$

This yields the equation for the image distance as:

Equation:

$$d_{\mathrm{i}} = \frac{fd_{\mathrm{o}}}{d_{\mathrm{o}} - f}.$$

Note that there is no inverting here.

The thin lens equations can be used to find the magnification m, since both d_{i} and d_{o} are known. Entering their values gives

Equation:

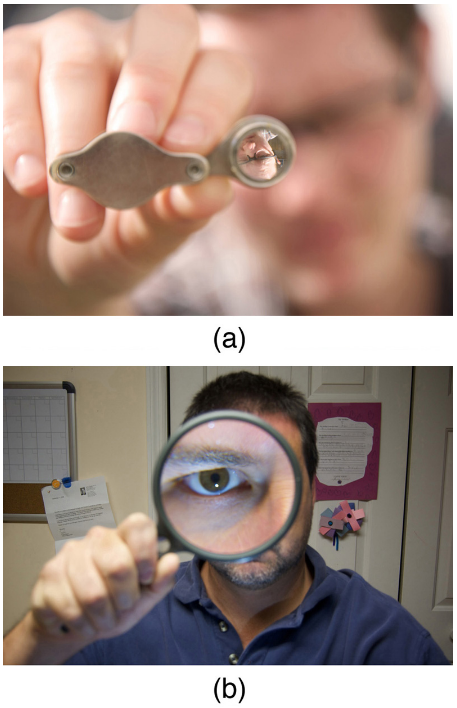
$$m = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{1.50\ \mathrm{m}}{0.750\ \mathrm{m}} = -2.00.$$

Discussion

Note that the minus sign causes the magnification to be negative when the image is inverted. Ray tracing and the use of the thin lens equations produce consistent results. The thin lens equations give the most precise results, being limited only by the accuracy of the given information. Ray tracing is limited by the accuracy with which you can draw, but it is highly useful both conceptually and visually.

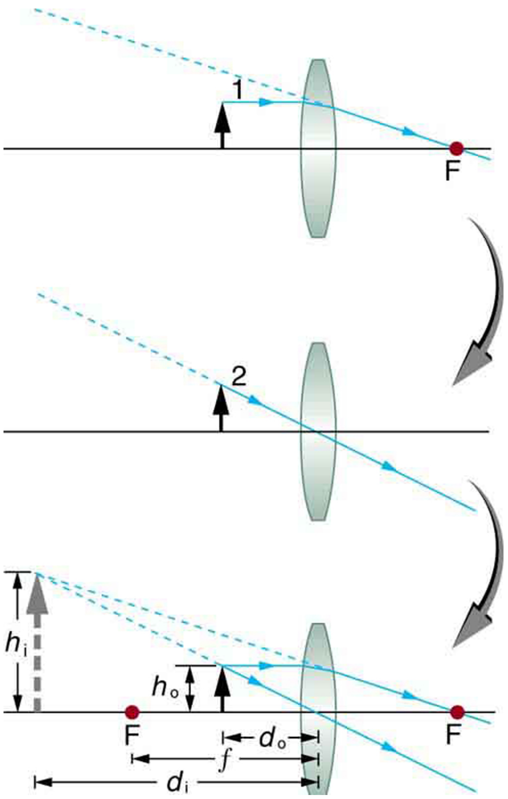
Real images, such as the one considered in the previous example, are formed by converging lenses whenever an object is farther from the lens than its focal length. This is true for movie projectors, cameras, and the eye. We shall refer to these as *case 1* images. A case 1 image is formed when $d_{\mathrm{o}} > f$ and f is positive, as in [link](a). (A summary of the three cases or types of image formation appears at the end of this section.)

A different type of image is formed when an object, such as a person's face, is held close to a convex lens. The image is upright and larger than the object, as seen in [link](b), and so the lens is called a magnifier. If you slowly pull the magnifier away from the face, you will see that the magnification steadily increases until the image begins to blur. Pulling the magnifier even farther away produces an inverted image as seen in [link](a). The distance at which the image blurs, and beyond which it inverts, is the focal length of the lens. To use a convex lens as a magnifier, the object must be closer to the converging lens than its focal length. This is called a *case 2* image. A case 2 image is formed when $d_\mathrm{o} < f$ and f is positive.



(a) When a converging lens is held farther away from the face than the lens’s focal length, an inverted image is formed. This is a case 1 image. Note that the image is in focus but the face is not, because the image is much closer to the camera taking this photograph than the face. (credit: DaMongMan, Flickr) (b) A magnified image of a face is produced by placing it closer to the converging lens than its focal length. This is a case 2 image. (credit: Casey Fleser, Flickr)

[link] uses ray tracing to show how an image is formed when an object is held closer to a converging lens than its focal length. Rays coming from a common point on the object continue to diverge after passing through the lens, but all appear to originate from a point at the location of the image. The image is on the same side of the lens as the object and is farther away from the lens than the object. This image, like all case 2 images, cannot be projected and, hence, is called a **virtual image**. Light rays only appear to originate at a virtual image; they do not actually pass through that location in space. A screen placed at the location of a virtual image will receive only diffuse light from the object, not focused rays from the lens. Additionally, a screen placed on the opposite side of the lens will receive rays that are still diverging, and so no image will be projected on it. We can see the magnified image with our eyes, because the lens of the eye converges the rays into a real image projected on our retina. Finally, we note that a virtual image is upright and larger than the object, meaning that the magnification is positive and greater than 1.



Ray tracing predicts the image location and size for an object held closer to a converging lens than its focal length. Ray 1 enters parallel to the axis and exits through the focal point on the opposite side, while ray 2 passes through the center of the lens without changing path. The two rays continue to diverge on the other side of the lens, but both appear to come from a common point, locating the upright, magnified,

virtual image. This is a
case 2 image.

Note:
Virtual Image
An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Example:
Image Produced by a Magnifying Glass
Suppose the book page in [link] (a) is held 7.50 cm from a convex lens of focal length 10.0 cm, such as a typical magnifying glass might have. What magnification is produced?

Strategy and Concept
We are given that $d_{\text{o}} = 7.50$ cm and $f = 10.0$ cm, so we have a situation where the object is placed closer to the lens than its focal length. We therefore expect to get a case 2 virtual image with a positive magnification that is greater than 1. Ray tracing produces an image like that shown in [link], but we will use the thin lens equations to get numerical solutions in this example.

Solution
To find the magnification m, we try to use magnification equation, $m = –d_{\text{i}}/d_{\text{o}}$. We do not have a value for d_{i}, so that we must first find the location of the image using lens equation. (The procedure is the same as followed in the preceding example, where d_{o} and f were known.) Rearranging the magnification equation to isolate d_{i} gives

Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{f} - \frac{1}{d_{\text{o}}}.$$

Entering known values, we obtain a value for $1/d_{\text{i}}$:

Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{10.0\text{ cm}} - \frac{1}{7.50\text{ cm}} = \frac{-0.0333}{\text{cm}}.$$

This must be inverted to find d_{i}:
Equation:

$$d_{\text{i}} = -\frac{\text{cm}}{0.0333} = -30.0\text{ cm}.$$

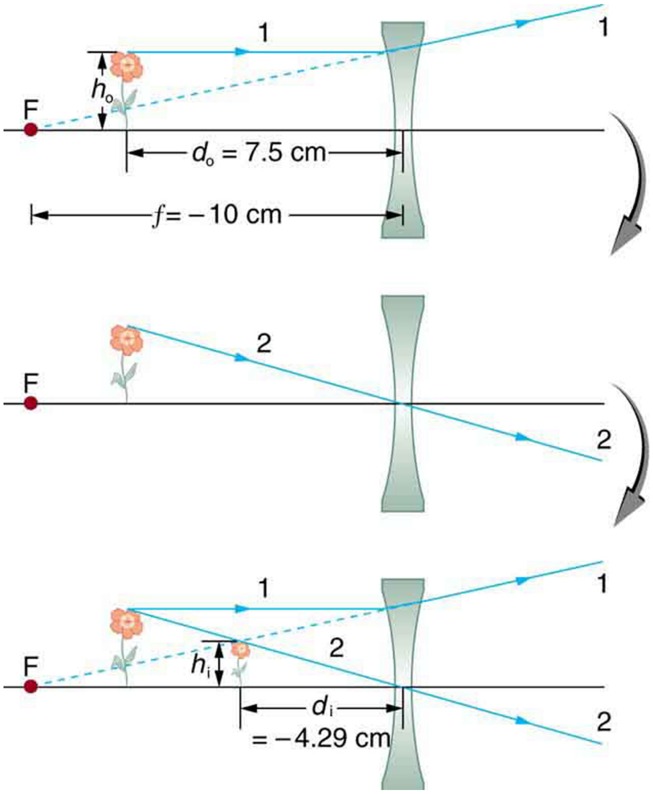
Now the thin lens equation can be used to find the magnification m, since both d_{i} and d_{o} are known. Entering their values gives
Equation:

$$m = -\frac{d_{\text{i}}}{d_{\text{o}}} = -\frac{-30.0\text{ cm}}{7.50\text{ cm}} = 4.00.$$

Discussion
A number of results in this example are true of all case 2 images, as well as being consistent with [link]. Magnification is indeed positive (as predicted), meaning the image is upright. The magnification is also greater than 1, meaning that the image is larger than the object—in this case, by a factor of 4. Note that the image distance is negative. This means the image is on the same side of the lens as the object. Thus the image cannot be projected and is virtual. (Negative values of d_{i} occur for virtual images.) The image is farther from the lens than the object, since the image distance is greater in magnitude than the object distance. The location of the image is not obvious when you look through a magnifier. In fact, since the image is bigger than the object, you may think the image is closer than the object. But the image is farther away, a fact that is useful in correcting farsightedness, as we shall see in a later section.

A third type of image is formed by a diverging or concave lens. Try looking through eyeglasses meant to correct nearsightedness. (See [link].) You will see an image that is upright but smaller than the object. This means that the magnification is positive but less than 1. The ray diagram in [link] shows that the image is on the same side of the lens as the object and, hence, cannot be projected—it is a virtual image. Note that the image is closer to the lens than the object. This is a *case 3* image, formed for any object by a negative focal length or diverging lens.

A car viewed through a concave or diverging lens looks upright. This is a case 3 image. (credit: Daniel Oines, Flickr)



Ray tracing predicts the image location and size for a concave or diverging lens. Ray 1 enters parallel to the axis and is bent so that it appears to originate from the focal point. Ray 2 passes through the center of the lens without changing path. The two rays appear to come from a common point, locating the upright image. This is a case 3 image, which is closer to the lens than the object and smaller in height.

Example:

Image Produced by a Concave Lens

Suppose an object such as a book page is held 7.50 cm from a concave lens of focal length –10.0 cm. Such a lens could be used in eyeglasses to correct pronounced nearsightedness. What magnification is produced?

Strategy and Concept

This example is identical to the preceding one, except that the focal length is negative for a concave or diverging lens. The method of solution is thus the same, but the results are different in important ways.

Solution

To find the magnification m, we must first find the image distance d_{i} using thin lens equation

Equation:

$$\frac{1}{d_{\text{i}}}=\frac{1}{f}-\frac{1}{d_{\text{o}}},$$

or its alternative rearrangement

Equation:

$$d_i=\frac{fd_{\text{o}}}{d_{\text{o}}-f}.$$

We are given that $f=$ –10.0 cm and $d_{\text{o}}=7.50$ cm. Entering these yields a value for $1/d_{\text{i}}$:

Equation:

$$\frac{1}{d_{\text{i}}}=\frac{1}{-10.0\text{ cm}}-\frac{1}{7.50\text{ cm}}=\frac{-0.2333}{\text{cm}}.$$

This must be inverted to find d_{i}:

Equation:

$$d_{\text{i}}=-\frac{\text{cm}}{0.2333}=-4.29\text{ cm}.$$

Or

Equation:

$$d_{\mathrm{i}} = \frac{(7.5)(-10)}{(7.5-(-10))} = -75/17.5 = -4.29\text{ cm}.$$

Now the magnification equation can be used to find the magnification m, since both d_{i} and d_{o} are known. Entering their values gives

Equation:

$$m = -\frac{d_{\mathrm{i}}}{d_{\mathrm{o}}} = -\frac{-4.29\text{ cm}}{7.50\text{ cm}} = 0.571.$$

Discussion

A number of results in this example are true of all case 3 images, as well as being consistent with [link]. Magnification is positive (as predicted), meaning the image is upright. The magnification is also less than 1, meaning the image is smaller than the object—in this case, a little over half its size. The image distance is negative, meaning the image is on the same side of the lens as the object. (The image is virtual.) The image is closer to the lens than the object, since the image distance is smaller in magnitude than the object distance. The location of the image is not obvious when you look through a concave lens. In fact, since the image is smaller than the object, you may think it is farther away. But the image is closer than the object, a fact that is useful in correcting nearsightedness, as we shall see in a later section.

[link] summarizes the three types of images formed by single thin lenses. These are referred to as case 1, 2, and 3 images. Convex (converging) lenses can form either real or virtual images (cases 1 and 2, respectively), whereas concave (diverging) lenses can form only virtual images (always case 3). Real images are always inverted, but they can be either larger or smaller than the object. For example, a slide projector forms an image larger than the slide, whereas a camera makes an image smaller than the object being photographed. Virtual images are always upright and cannot be projected. Virtual images are larger than the object only in case 2, where a convex lens is used. The virtual image produced by a concave lens is

always smaller than the object—a case 3 image. We can see and photograph virtual images only by using an additional lens to form a real image.

Type	**Formed when**	**Image type**	***d*~i~**	***m***
Case 1	f positive, $d_o > f$	real	positive	negative
Case 2	f positive, $d_o < f$	virtual	negative	positive $m > 1$
Case 3	f negative	virtual	negative	positive $m < 1$

Three Types of Images Formed By Thin Lenses

In Image Formation by Mirrors, we shall see that mirrors can form exactly the same types of images as lenses.

Note:

Take-Home Experiment: Concentrating Sunlight

Find several lenses and determine whether they are converging or diverging. In general those that are thicker near the edges are diverging and those that are thicker near the center are converging. On a bright sunny day take the converging lenses outside and try focusing the sunlight onto a piece of paper. Determine the focal lengths of the lenses. Be careful because the paper may start to burn, depending on the type of lens you have selected.

Problem-Solving Strategies for Lenses

Step 1. Examine the situation to determine that image formation by a lens is involved.

Step 2. Determine whether ray tracing, the thin lens equations, or both are to be employed. A sketch is very useful even if ray tracing is not specifically required by the problem. Write symbols and values on the sketch.

Step 3. Identify exactly what needs to be determined in the problem (identify the unknowns).

Step 4. Make alist of what is given or can be inferred from the problem as stated (identify the knowns). It is helpful to determine whether the situation involves a case 1, 2, or 3 image. While these are just names for types of images, they have certain characteristics (given in [link]) that can be of great use in solving problems.

Step 5. If ray tracing is required, use the ray tracing rules listed near the beginning of this section.

Step 6. Most quantitative problems require the use of the thin lens equations. These are solved in the usual manner by substituting knowns and solving for unknowns. Several worked examples serve as guides.

Step 7. Check to see if the answer is reasonable: Does it make sense? If you have identified the type of image (case 1, 2, or 3), you should assess whether your answer is consistent with the type of image, magnification, and so on.

Note:
Misconception Alert
We do not realize that light rays are coming from every part of the object, passing through every part of the lens, and all can be used to form the final image.
We generally feel the entire lens, or mirror, is needed to form an image. Actually, half a lens will form the same, though a fainter, image.

Section Summary

- Light rays entering a converging lens parallel to its axis cross one another at a single point on the opposite side.
- For a converging lens, the focal point is the point at which converging light rays cross; for a diverging lens, the focal point is the point from which diverging light rays appear to originate.
- The distance from the center of the lens to its focal point is called the focal length f.
- Power P of a lens is defined to be the inverse of its focal length, $P=\frac{1}{f}$.
- A lens that causes the light rays to bend away from its axis is called a diverging lens.
- Ray tracing is the technique of graphically determining the paths that light rays take.
- The image in which light rays from one point on the object actually cross at the location of the image and can be projected onto a screen, a piece of film, or the retina of an eye is called a real image.
- Thin lens equations are $\frac{1}{d_{\text{o}}}+\frac{1}{d_{\text{i}}}=\frac{1}{f}$ and $\frac{h_{\text{i}}}{h_{\text{o}}}=-\frac{d_{\text{i}}}{d_{\text{o}}}=m$ (magnification).

- The distance of the image from the center of the lens is called image distance.
- An image that is on the same side of the lens as the object and cannot be projected on a screen is called a virtual image.

Conceptual Questions

Exercise:

Problem:

It can be argued that a flat piece of glass, such as in a window, is like a lens with an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?

Exercise:

Problem:

You can often see a reflection when looking at a sheet of glass, particularly if it is darker on the other side. Explain why you can often see a double image in such circumstances.

Exercise:

Problem:

When you focus a camera, you adjust the distance of the lens from the film. If the camera lens acts like a thin lens, why can it not be a fixed distance from the film for both near and distant objects?

Exercise:

Problem:

A thin lens has two focal points, one on either side, at equal distances from its center, and should behave the same for light entering from either side. Look through your eyeglasses (or those of a friend) backward and forward and comment on whether they are thin lenses.

Exercise:

Problem:

Will the focal length of a lens change when it is submerged in water? Explain.

Problems & Exercises

Exercise:

Problem:

What is the power in diopters of a camera lens that has a 50.0 mm focal length?

Exercise:

Problem:

Your camera’s zoom lens has an adjustable focal length ranging from 80.0 to 200 mm. What is its range of powers?

Solution:

$5.00 \text{ to } 12.5 \text{ D}$

Exercise:

Problem:

What is the focal length of 1.75 D reading glasses found on the rack in a pharmacy?

Exercise:

Problem:

You note that your prescription for new eyeglasses is –4.50 D. What will their focal length be?

Solution:

-0.222 m

Exercise:

Problem:

How far from the lens must the film in a camera be, if the lens has a 35.0 mm focal length and is being used to photograph a flower 75.0 cm away? Explicitly show how you follow the steps in the Problem-Solving Strategy for lenses.

Exercise:

Problem:

A certain slide projector has a 100 mm focal length lens. (a) How far away is the screen, if a slide is placed 103 mm from the lens and produces a sharp image? (b) If the slide is 24.0 by 36.0 mm, what are the dimensions of the image? Explicitly show how you follow the steps in the Problem-Solving Strategy for lenses.

Solution:

(a) 3.43 m

(b) 0.800 by 1.20 m

Exercise:

Problem:

A doctor examines a mole with a 15.0 cm focal length magnifying glass held 13.5 cm from the mole (a) Where is the image? (b) What is its magnification? (c) How big is the image of a 5.00 mm diameter mole?

Solution:

(a) -1.35 m (on the object side of the lens).

(b) $+10.0$

(c) 5.00 cm

Exercise:

Problem:

How far from a piece of paper must you hold your father’s 2.25 D reading glasses to try to burn a hole in the paper with sunlight?

Solution:

44.4 cm

Exercise:

Problem:

A camera with a 50.0 mm focal length lens is being used to photograph a person standing 3.00 m away. (a) How far from the lens must the film be? (b) If the film is 36.0 mm high, what fraction of a 1.75 m tall person will fit on it? (c) Discuss how reasonable this seems, based on your experience in taking or posing for photographs.

Exercise:

Problem:

A camera lens used for taking close-up photographs has a focal length of 22.0 mm. The farthest it can be placed from the film is 33.0 mm. (a) What is the closest object that can be photographed? (b) What is the magnification of this closest object?

Solution:

(a) 6.60 cm

(b) –0.333

Exercise:

Problem:

Suppose your 50.0 mm focal length camera lens is 51.0 mm away from the film in the camera. (a) How far away is an object that is in focus? (b) What is the height of the object if its image is 2.00 cm high?

Exercise:

Problem:

(a) What is the focal length of a magnifying glass that produces a magnification of 3.00 when held 5.00 cm from an object, such as a rare coin? (b) Calculate the power of the magnifier in diopters. (c) Discuss how this power compares to those for store-bought reading glasses (typically 1.0 to 4.0 D). Is the magnifier's power greater, and should it be?

Solution:

(a) $+7.50\text{ cm}$

(b) 13.3 D

(c) Much greater

Exercise:

Problem:

What magnification will be produced by a lens of power –4.00 D (such as might be used to correct myopia) if an object is held 25.0 cm away?

Exercise:

Problem:

In [link], the magnification of a book held 7.50 cm from a 10.0 cm focal length lens was found to be 3.00. (a) Find the magnification for the book when it is held 8.50 cm from the magnifier. (b) Do the same for when it is held 9.50 cm from the magnifier. (c) Comment on the trend in m as the object distance increases as in these two calculations.

Solution:

(a) +6.67

(b) +20.0

(c) The magnification increases without limit (to infinity) as the object distance increases to the limit of the focal distance.

Exercise:

Problem:

Suppose a 200 mm focal length telephoto lens is being used to photograph mountains 10.0 km away. (a) Where is the image? (b) What is the height of the image of a 1000 m high cliff on one of the mountains?

Exercise:

Problem:

A camera with a 100 mm focal length lens is used to photograph the sun and moon. What is the height of the image of the sun on the film, given the sun is 1.40×10^6 km in diameter and is 1.50×10^8 km away?

Solution:

-0.933 mm

Exercise:

Problem:

Combine thin lens equations to show that the magnification for a thin lens is determined by its focal length and the object distance and is given by $m = f/(f - d_{\text{o}})$.

Glossary

converging lens
a convex lens in which light rays that enter it parallel to its axis converge at a single point on the opposite side

diverging lens
a concave lens in which light rays that enter it parallel to its axis bend away (diverge) from its axis

focal point
for a converging lens or mirror, the point at which converging light rays cross; for a diverging lens or mirror, the point from which diverging light rays appear to originate

focal length
distance from the center of a lens or curved mirror to its focal point

magnification
ratio of image height to object height

power
inverse of focal length

real image
image that can be projected

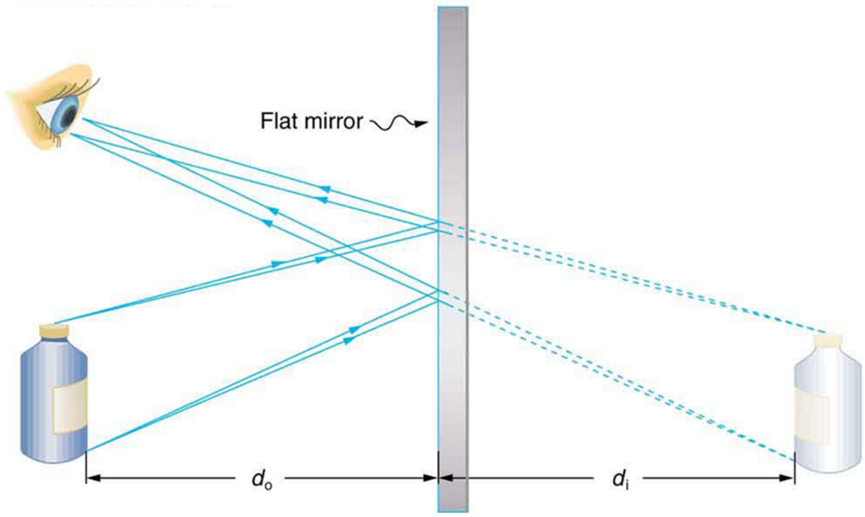
virtual image
image that cannot be projected

Image Formation by Mirrors

- Illustrate image formation in a flat mirror.
- Explain with ray diagrams the formation of an image using spherical mirrors.
- Determine focal length and magnification given radius of curvature, distance of object and image.

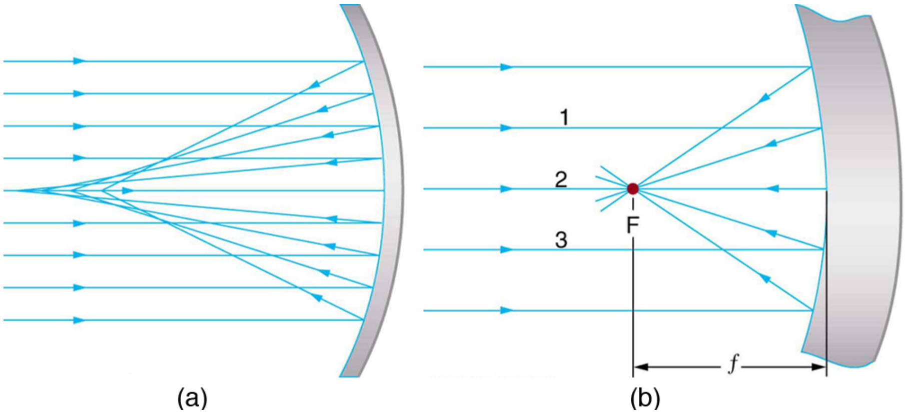
We only have to look as far as the nearest bathroom to find an example of an image formed by a mirror. Images in flat mirrors are the same size as the object and are located behind the mirror. Like lenses, mirrors can form a variety of images. For example, dental mirrors may produce a magnified image, just as makeup mirrors do. Security mirrors in shops, on the other hand, form images that are smaller than the object. We will use the law of reflection to understand how mirrors form images, and we will find that mirror images are analogous to those formed by lenses.

[link] helps illustrate how a flat mirror forms an image. Two rays are shown emerging from the same point, striking the mirror, and being reflected into the observer’s eye. The rays can diverge slightly, and both still get into the eye. If the rays are extrapolated backward, they seem to originate from a common point behind the mirror, locating the image. (The paths of the reflected rays into the eye are the same as if they had come directly from that point behind the mirror.) Using the law of reflection—the angle of reflection equals the angle of incidence—we can see that the image and object are the same distance from the mirror. This is a virtual image, since it cannot be projected—the rays only appear to originate from a common point behind the mirror. Obviously, if you walk behind the mirror, you cannot see the image, since the rays do not go there. But in front of the mirror, the rays behave exactly as if they had come from behind the mirror, so that is where the image is situated.



Two sets of rays from common points on an object are reflected by a flat mirror into the eye of an observer. The reflected rays seem to originate from behind the mirror, locating the virtual image.

Now let us consider the focal length of a mirror—for example, the concave spherical mirrors in [link]. Rays of light that strike the surface follow the law of reflection. For a mirror that is large compared with its radius of curvature, as in [link](a), we see that the reflected rays do not cross at the same point, and the mirror does not have a well-defined focal point. If the mirror had the shape of a parabola, the rays would all cross at a single point, and the mirror would have a well-defined focal point. But parabolic mirrors are much more expensive to make than spherical mirrors. The solution is to use a mirror that is small compared with its radius of curvature, as shown in [link](b). (This is the mirror equivalent of the thin lens approximation.) To a very good approximation, this mirror has a well-defined focal point at F that is the focal distance f from the center of the mirror. The focal length f of a concave mirror is positive, since it is a converging mirror.



(a) Parallel rays reflected from a large spherical mirror do not all cross at a common point. (b) If a spherical mirror is small compared with its radius of curvature, parallel rays are focused to a common point. The distance of the focal point from the center of the mirror is its focal length f. Since this mirror is converging, it has a positive focal length.

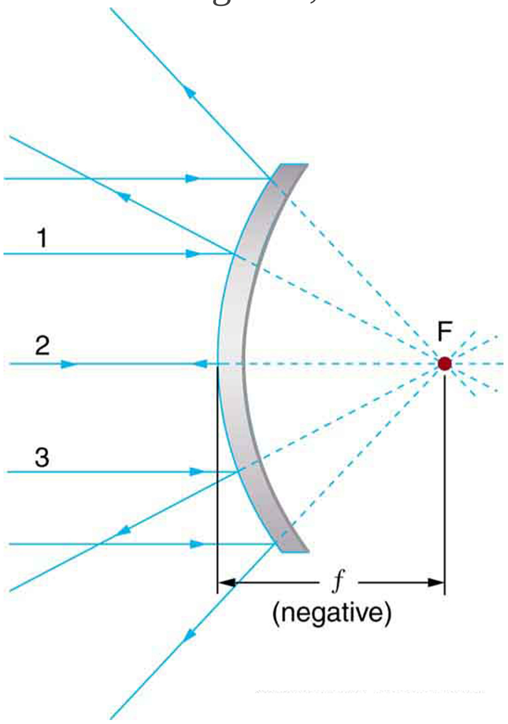
Just as for lenses, the shorter the focal length, the more powerful the mirror; thus, $P = 1/f$ for a mirror, too. A more strongly curved mirror has a shorter focal length and a greater power. Using the law of reflection and some simple trigonometry, it can be shown that the focal length is half the radius of curvature, or

Equation:

$$f = \frac{R}{2},$$

where R is the radius of curvature of a spherical mirror. The smaller the radius of curvature, the smaller the focal length and, thus, the more powerful the mirror.

The convex mirror shown in [link] also has a focal point. Parallel rays of light reflected from the mirror seem to originate from the point F at the focal distance f behind the mirror. The focal length and power of a convex mirror are negative, since it is a diverging mirror.



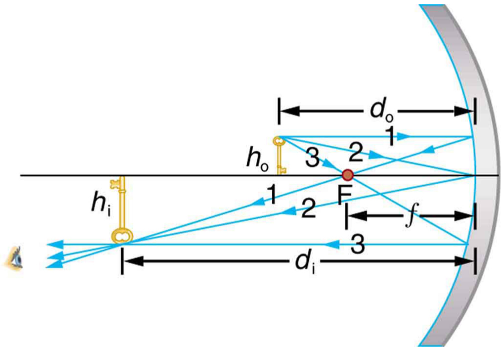
Parallel rays of light reflected from a convex spherical mirror (small in size compared with its radius of curvature) seem to originate from a well-defined focal point at the focal distance f behind the mirror. Convex mirrors diverge light rays and, thus, have a negative focal length.

Ray tracing is as useful for mirrors as for lenses. The rules for ray tracing for mirrors are based on the illustrations just discussed:

1. A ray approaching a concave converging mirror parallel to its axis is reflected through the focal point F of the mirror on the same side. (See rays 1 and 3 in [link](b).)
2. A ray approaching a convex diverging mirror parallel to its axis is reflected so that it seems to come from the focal point F behind the mirror. (See rays 1 and 3 in [link].)
3. Any ray striking the center of a mirror is followed by applying the law of reflection; it makes the same angle with the axis when leaving as when approaching. (See ray 2 in [link].)
4. A ray approaching a concave converging mirror through its focal point is reflected parallel to its axis. (The reverse of rays 1 and 3 in [link].)
5. A ray approaching a convex diverging mirror by heading toward its focal point on the opposite side is reflected parallel to the axis. (The reverse of rays 1 and 3 in [link].)

We will use ray tracing to illustrate how images are formed by mirrors, and we can use ray tracing quantitatively to obtain numerical information. But since we assume each mirror is small compared with its radius of curvature, we can use the thin lens equations for mirrors just as we did for lenses.

Consider the situation shown in [link], concave spherical mirror reflection, in which an object is placed farther from a concave (converging) mirror than its focal length. That is, f is positive and $d_o > f$, so that we may expect an image similar to the case 1 real image formed by a converging lens. Ray tracing in [link] shows that the rays from a common point on the object all cross at a point on the same side of the mirror as the object. Thus a real image can be projected onto a screen placed at this location. The image distance is positive, and the image is inverted, so its magnification is negative. This is a *case 1 image for mirrors*. It differs from the case 1 image for lenses only in that the image is on the same side of the mirror as the object. It is otherwise identical.



A case 1 image for a mirror. An object is farther from the converging mirror than its focal length. Rays from a common point on the object are traced using the rules in the text. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 goes through the focal point on the way toward the mirror. All three rays cross at the same point after being reflected, locating the inverted real image. Although three rays are shown, only two of the three are needed to locate the image and determine its height.

Example:

A Concave Reflector

Electric room heaters use a concave mirror to reflect infrared (IR) radiation from hot coils. Note that IR follows the same law of reflection as visible light. Given that the mirror has a radius of curvature of 50.0 cm and produces an image of the coils 3.00 m away from the mirror, where are the coils?

Strategy and Concept

We are given that the concave mirror projects a real image of the coils at an image distance $d_\text{i} = 3.00\text{ m}$. The coils are the object, and we are asked to find their location—that is, to find the object distance d_o. We are also given the radius of curvature of the mirror, so that its focal length is $f = R/2 = 25.0\text{ cm}$ (positive since the mirror is concave or converging). Assuming the mirror is small compared with its radius of curvature, we can use the thin lens equations, to solve this problem.

Solution

Since d_i and f are known, thin lens equation can be used to find d_o:

Equation:

$$\frac{1}{d_\text{o}} + \frac{1}{d_\text{i}} = \frac{1}{f}.$$

Rearranging to isolate d_o gives

Equation:

$$\frac{1}{d_\text{o}} = \frac{1}{f} - \frac{1}{d_\text{i}}.$$

Entering known quantities gives a value for $1/d_\text{o}$:

Equation:

$$\frac{1}{d_\text{o}} = \frac{1}{0.250\text{ m}} - \frac{1}{3.00\text{ m}} = \frac{3.667}{\text{m}}.$$

This must be inverted to find d_o:

Equation:

$$d_\text{o} = \frac{1\text{ m}}{3.667} = 27.3\text{ cm}.$$

Discussion

Note that the object (the filament) is farther from the mirror than the mirror’s focal length. This is a case 1 image ($d_\text{o} > f$ and f positive), consistent with the fact that a real image is formed. You will get the most concentrated thermal energy directly in front of the mirror and 3.00 m away from it. Generally, this is not desirable, since it could cause burns. Usually, you want the rays to emerge parallel, and this is accomplished by having the filament at the focal point of the mirror.

Note that the filament here is not much farther from the mirror than its focal length and that the image produced is considerably farther away. This is exactly analogous to a slide projector. Placing a slide only slightly farther away from the projector lens than its focal length produces an image significantly farther away. As the object gets closer to the focal distance, the image gets farther away. In fact, as the object distance approaches the focal length, the image distance approaches infinity and the rays are sent out parallel to one another.

Example:

Solar Electric Generating System

One of the solar technologies used today for generating electricity is a device (called a parabolic trough or concentrating collector) that concentrates the sunlight onto a blackened pipe that contains a fluid. This heated fluid is pumped to a heat exchanger, where its heat energy is transferred to another system that is used to generate steam—and so generate electricity through a conventional steam cycle. [link] shows such a working system in southern California. Concave mirrors are used to concentrate the sunlight onto the pipe. The mirror has the approximate shape of a section of a cylinder. For the problem, assume that the mirror is exactly one-quarter of a full cylinder.

a. If we wish to place the fluid-carrying pipe 40.0 cm from the concave mirror at the mirror’s focal point, what will be the radius of curvature of the mirror?
b. Per meter of pipe, what will be the amount of sunlight concentrated onto the pipe, assuming the insolation (incident solar radiation) is

$0.900\ \text{kW/m}^2$?

c. If the fluid-carrying pipe has a 2.00-cm diameter, what will be the temperature increase of the fluid per meter of pipe over a period of one minute? Assume all the solar radiation incident on the reflector is absorbed by the pipe, and that the fluid is mineral oil.

Strategy

To solve an *Integrated Concept Problem* we must first identify the physical principles involved. Part (a) is related to the current topic. Part (b) involves a little math, primarily geometry. Part (c) requires an understanding of heat and density.

Solution to (a)

To a good approximation for a concave or semi-spherical surface, the point where the parallel rays from the sun converge will be at the focal point, so $R = 2f = 80.0\ \text{cm}$.

Solution to (b)

The insolation is $900\ \text{W/m}^2$. We must find the cross-sectional area A of the concave mirror, since the power delivered is $900\ \text{W/m}^2 \times \text{A}$. The mirror in this case is a quarter-section of a cylinder, so the area for a length L of the mirror is $\text{A} = \frac{1}{4}(2\pi R)\text{L}$. The area for a length of $1.00\ \text{m}$ is then

Equation:

$$\text{A} = \frac{\pi}{2}R(1.00\ \text{m}) = \frac{(3.14)}{2}(0.800\ \text{m})(1.00\ \text{m}) = 1.26\ \text{m}^2.$$

The insolation on the 1.00-m length of pipe is then

Equation:

$$\left(9.00 \times 10^2 \frac{\text{W}}{\text{m}^2}\right)\left(1.26\ \text{m}^2\right) = 1130\ \text{W}.$$

Solution to (c)

The increase in temperature is given by $Q = mc\Delta T$. The mass m of the mineral oil in the one-meter section of pipe is

Equation:

$$\begin{aligned} m &= \rho V = \rho\pi\left(\frac{d}{2}\right)^2(1.00\text{ m}) \\ &= \left(8.00\times10^2\text{ kg/m}^3\right)(3.14)(0.0100\text{ m})^2(1.00\text{ m}) \\ &= 0.251\text{ kg.} \end{aligned}$$

Therefore, the increase in temperature in one minute is

Equation:

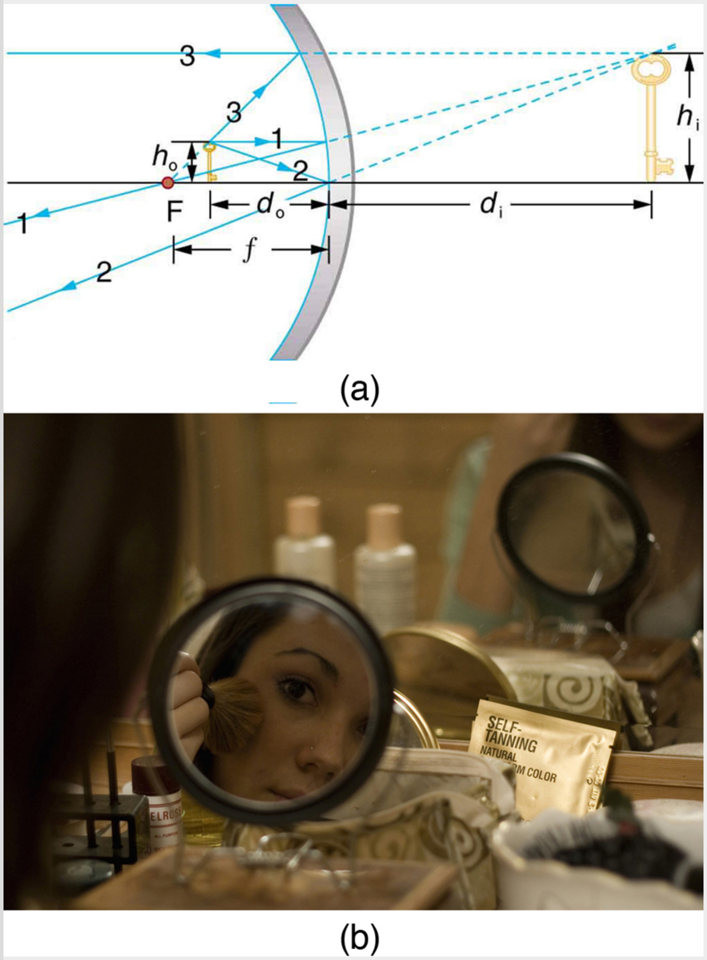
$$\begin{aligned} \Delta T &= Q/mc \\ &= \frac{(1130\text{ W})(60.0\text{ s})}{(0.251\text{ kg})(1670\text{ J}\cdot\text{kg/}^\circ\text{C})} \\ &= 162^\circ\text{C.} \end{aligned}$$

Discussion for (c)

An array of such pipes in the California desert can provide a thermal output of 250 MW on a sunny day, with fluids reaching temperatures as high as 400ºC. We are considering only one meter of pipe here, and ignoring heat losses along the pipe.

Parabolic trough collectors are used to generate electricity in southern California. (credit: kjkolb, Wikimedia Commons)

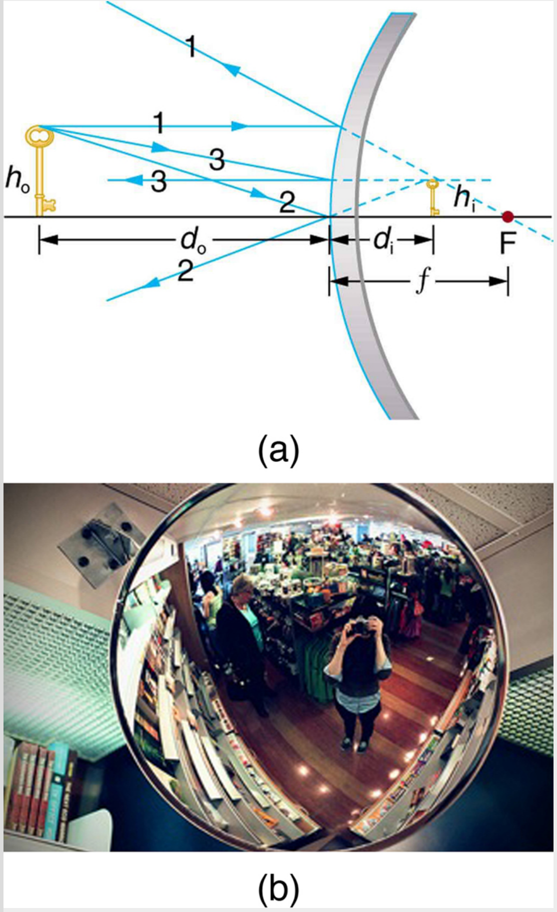
What happens if an object is closer to a concave mirror than its focal length? This is analogous to a case 2 image for lenses ($d_o < f$ and f positive), which is a magnifier. In fact, this is how makeup mirrors act as magnifiers. [link](a) uses ray tracing to locate the image of an object placed close to a concave mirror. Rays from a common point on the object are reflected in such a manner that they appear to be coming from behind the mirror, meaning that the image is virtual and cannot be projected. As with a magnifying glass, the image is upright and larger than the object. This is a *case 2 image for mirrors* and is exactly analogous to that for lenses.



(a) Case 2 images for mirrors are formed when a converging mirror has an object closer to it than its focal length. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches the mirror as if it came from the focal point. (b) A magnifying mirror showing the reflection. (credit: Mike Melrose, Flickr)

All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be larger than the object. (b) Makeup mirrors are perhaps the most common use of a concave mirror to produce a larger, upright image.

A convex mirror is a diverging mirror (f is negative) and forms only one type of image. It is a *case 3* image—one that is upright and smaller than the object, just as for diverging lenses. [link](a) uses ray tracing to illustrate the location and size of the case 3 image for mirrors. Since the image is behind the mirror, it cannot be projected and is thus a virtual image. It is also seen to be smaller than the object.



(a)

(b)

Case 3 images for mirrors are formed by any convex mirror. Ray 1 approaches parallel to the axis, ray 2 strikes the center of the mirror, and ray 3 approaches toward the focal point. All three rays appear to originate from the same point after being reflected, locating the upright virtual image behind the mirror and showing it to be smaller than the object. (b) Security mirrors are convex, producing a smaller, upright image. Because the image is smaller, a larger area is imaged compared to what would be observed for a flat mirror (and hence security is improved). (credit: Laura D'Alessandro, Flickr)

Example:

Image in a Convex Mirror

A keratometer is a device used to measure the curvature of the cornea, particularly for fitting contact lenses. Light is reflected from the cornea, which acts like a convex mirror, and the keratometer measures the magnification of the image. The smaller the magnification, the smaller the radius of curvature of the cornea. If the light source is 12.0 cm from the cornea and the image's magnification is 0.0320, what is the cornea's radius of curvature?

Strategy

If we can find the focal length of the convex mirror formed by the cornea, we can find its radius of curvature (the radius of curvature is twice the focal length of a spherical mirror). We are given that the object distance is

$d_{\text{o}} = 12.0\text{ cm}$ and that $m = 0.0320$. We first solve for the image distance d_{i}, and then for f.

Solution

$m = –d_{\text{i}}/d_{\text{o}}$. Solving this expression for d_{i} gives

Equation:

$$d_{\text{i}} = -md_{\text{o}}.$$

Entering known values yields

Equation:

$$d_{\text{i}} =– (0.0320)(12.0\text{ cm}) = –0.384\text{ cm}.$$

Equation:

$$\frac{1}{f} = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$$

Substituting known values,

Equation:

$$\frac{1}{f} = \frac{1}{12.0\text{ cm}} + \frac{1}{-0.384\text{ cm}} = \frac{-2.52}{\text{cm}}.$$

This must be inverted to find f:

Equation:

$$f = \frac{\text{cm}}{–2.52} = –0.400\text{ cm}.$$

The radius of curvature is twice the focal length, so that

Equation:

$$R = 2 \mid f \mid = 0.800\text{ cm}.$$

Discussion

Although the focal length f of a convex mirror is defined to be negative, we take the absolute value to give us a positive value for R. The radius of curvature found here is reasonable for a cornea. The distance from cornea to retina in an adult eye is about 2.0 cm. In practice, many corneas are not spherical, complicating the job of fitting contact lenses. Note that the image distance here is negative, consistent with the fact that the image is behind the mirror, where it cannot be projected. In this section's Problems and Exercises, you will show that for a fixed object distance, the smaller the radius of curvature, the smaller the magnification.

The three types of images formed by mirrors (cases 1, 2, and 3) are exactly analogous to those formed by lenses, as summarized in the table at the end of Image Formation by Lenses. It is easiest to concentrate on only three types of images—then remember that concave mirrors act like convex lenses, whereas convex mirrors act like concave lenses.

Note:

Take-Home Experiment: Concave Mirrors Close to Home

Find a flashlight and identify the curved mirror used in it. Find another flashlight and shine the first flashlight onto the second one, which is turned off. Estimate the focal length of the mirror. You might try shining a flashlight on the curved mirror behind the headlight of a car, keeping the headlight switched off, and determine its focal length.

Problem-Solving Strategy for Mirrors

Step 1. Examine the situation to determine that image formation by a mirror is involved.

Step 2. Refer to the Problem-Solving Strategies for Lenses. The same strategies are valid for mirrors as for lenses with one qualification—use the ray tracing rules for mirrors listed earlier in this section.

Section Summary

- The characteristics of an image formed by a flat mirror are: (a) The image and object are the same distance from the mirror, (b) The image is a virtual image, and (c) The image is situated behind the mirror.

- Image length is half the radius of curvature.
  Equation:

$$f = \frac{R}{2}$$

- A convex mirror is a diverging mirror and forms only one type of image, namely a virtual image.

Conceptual Questions

Exercise:

Problem:

What are the differences between real and virtual images? How can you tell (by looking) whether an image formed by a single lens or mirror is real or virtual?

Exercise:

Problem:

Can you see a virtual image? Can you photograph one? Can one be projected onto a screen with additional lenses or mirrors? Explain your responses.

Exercise:

Problem:

Is it necessary to project a real image onto a screen for it to exist?

Exercise:

Problem:

At what distance is an image *always* located—at d_{o}, d_{i}, or f?

Exercise:

Problem:

Under what circumstances will an image be located at the focal point of a lens or mirror?

Exercise:

Problem:

What is meant by a negative magnification? What is meant by a magnification that is less than 1 in magnitude?

Exercise:

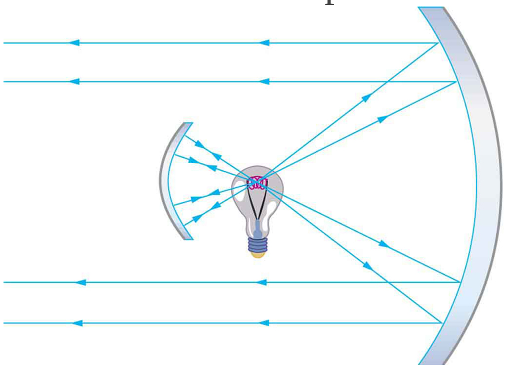
Problem:

Can a case 1 image be larger than the object even though its magnification is always negative? Explain.

Exercise:

Problem:

[link] shows a light bulb between two mirrors. One mirror produces a beam of light with parallel rays; the other keeps light from escaping without being put into the beam. Where is the filament of the light in relation to the focal point or radius of curvature of each mirror?



The two mirrors trap most of the bulb’s light and form a directional beam as in a headlight.

Exercise:

Problem:

Devise an arrangement of mirrors allowing you to see the back of your head. What is the minimum number of mirrors needed for this task?

Exercise:

Problem:

If you wish to see your entire body in a flat mirror (from head to toe), how tall should the mirror be? Does its size depend upon your distance away from the mirror? Provide a sketch.

Exercise:

Problem:

It can be argued that a flat mirror has an infinite focal length. If so, where does it form an image? That is, how are d_i and d_o related?

Exercise:

Problem:

Why are diverging mirrors often used for rear-view mirrors in vehicles? What is the main disadvantage of using such a mirror compared with a flat one?

Problems & Exercises

Exercise:

Problem:

What is the focal length of a makeup mirror that has a power of 1.50 D?

Solution:

+0.667 m

Exercise:

Problem:

Some telephoto cameras use a mirror rather than a lens. What radius of curvature mirror is needed to replace a 800 mm focal length telephoto lens?

Exercise:

Problem:

(a) Calculate the focal length of the mirror formed by the shiny back of a spoon that has a 3.00 cm radius of curvature. (b) What is its power in diopters?

Solution:

(a) -1.5×10^{-2} m

(b)–66.7 D

Exercise:

Problem:

Find the magnification of the heater element in [link]. Note that its large magnitude helps spread out the reflected energy.

Exercise:

Problem:

What is the focal length of a makeup mirror that produces a magnification of 1.50 when a person’s face is 12.0 cm away? Explicitly show how you follow the steps in the Problem-Solving Strategy for Mirrors.

Solution:

+0.360 m (concave)

Exercise:

Problem:

A shopper standing 3.00 m from a convex security mirror sees his image with a magnification of 0.250. (a) Where is his image? (b) What is the focal length of the mirror? (c) What is its radius of curvature? Explicitly show how you follow the steps in the Problem-Solving Strategy for Mirrors.

Exercise:

Problem:

An object 1.50 cm high is held 3.00 cm from a person’s cornea, and its reflected image is measured to be 0.167 cm high. (a) What is the magnification? (b) Where is the image? (c) Find the radius of curvature of the convex mirror formed by the cornea. (Note that this technique is used by optometrists to measure the curvature of the cornea for contact lens fitting. The instrument used is called a keratometer, or curve measurer.)

Solution:

(a) +0.111

(b) -0.334 cm (behind “mirror”)

(c) 0.752cm

Exercise:

Problem:

Ray tracing for a flat mirror shows that the image is located a distance behind the mirror equal to the distance of the object from the mirror. This is stated $d_\text{i} = -d_\text{o}$, since this is a negative image distance (it is a virtual image). (a) What is the focal length of a flat mirror? (b) What is its power?

Exercise:

Problem:

Show that for a flat mirror $h_{\text{i}} = h_{\text{o}}$, knowing that the image is a distance behind the mirror equal in magnitude to the distance of the object from the mirror.

Solution:
Equation:

$$m = \frac{h_{\text{i}}}{h_{\text{o}}} = -\frac{d_{\text{i}}}{d_{\text{o}}} = -\frac{-d_{\text{o}}}{d_{\text{o}}} = \frac{d_{\text{o}}}{d_{\text{o}}} = 1 \Rightarrow h_{\text{i}} = h_{\text{o}}$$

Exercise:

Problem:

Use the law of reflection to prove that the focal length of a mirror is half its radius of curvature. That is, prove that $f = R/2$. Note this is true for a spherical mirror only if its diameter is small compared with its radius of curvature.

Exercise:

Problem:

Referring to the electric room heater considered in the first example in this section, calculate the intensity of IR radiation in W/m^2 projected by the concave mirror on a person 3.00 m away. Assume that the heating element radiates 1500 W and has an area of $100\ \text{cm}^2$, and that half of the radiated power is reflected and focused by the mirror.

Solution:

$6.82\ \text{kW/m}^2$

Exercise:

Problem:

Consider a 250-W heat lamp fixed to the ceiling in a bathroom. If the filament in one light burns out then the remaining three still work. Construct a problem in which you determine the resistance of each filament in order to obtain a certain intensity projected on the bathroom floor. The ceiling is 3.0 m high. The problem will need to involve concave mirrors behind the filaments. Your instructor may wish to guide you on the level of complexity to consider in the electrical components.

Glossary

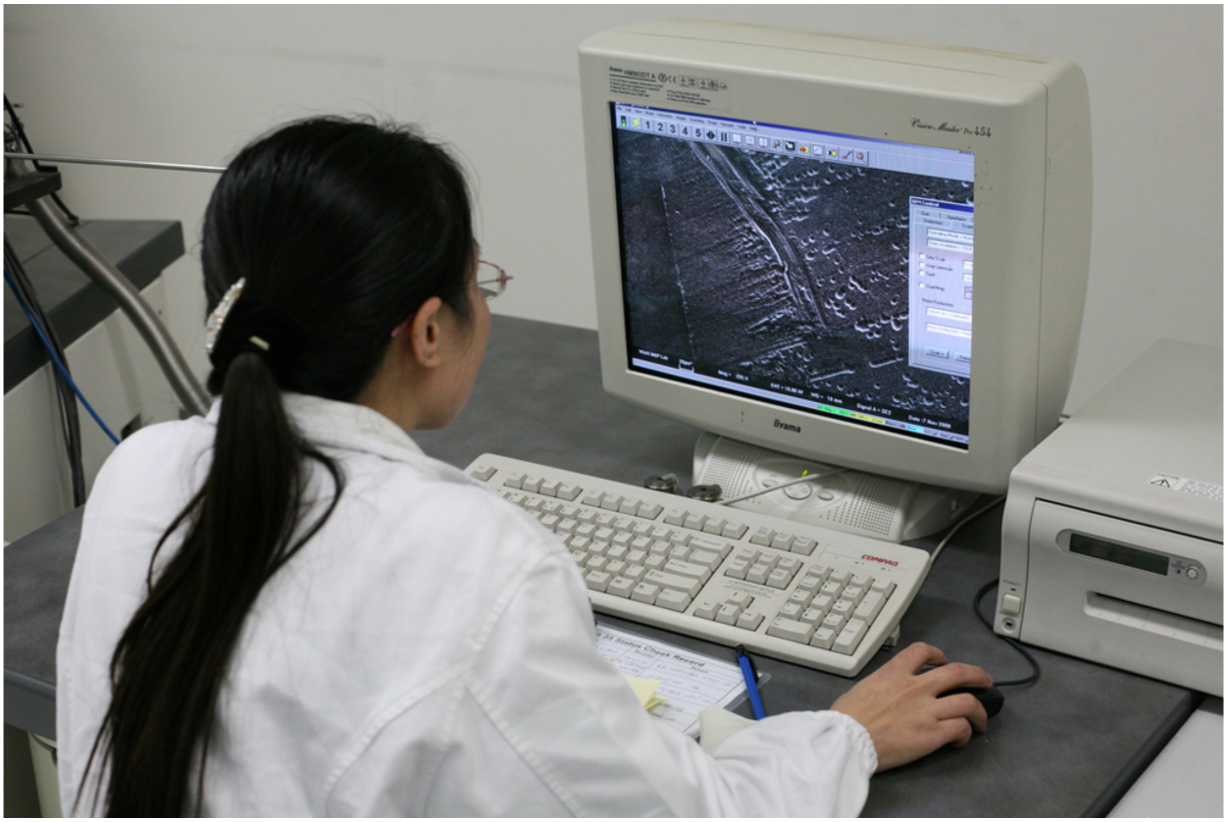
converging mirror
a concave mirror in which light rays that strike it parallel to its axis converge at one or more points along the axis

diverging mirror
a convex mirror in which light rays that strike it parallel to its axis bend away (diverge) from its axis

law of reflection
angle of reflection equals the angle of incidence

Introduction to Vision and Optical Instruments class="introduction"

A scientist examines minute details on the surface of a disk drive at a magnification of 100,000 times. The image was produced using an electron microscope. (credit: Robert Scoble)



Explore how the image on the computer screen is formed. How is the image formation on the computer screen different from the image formation in your eye as you look down the microscope? How can videos of living cell processes be taken for viewing later on, and by many different people?

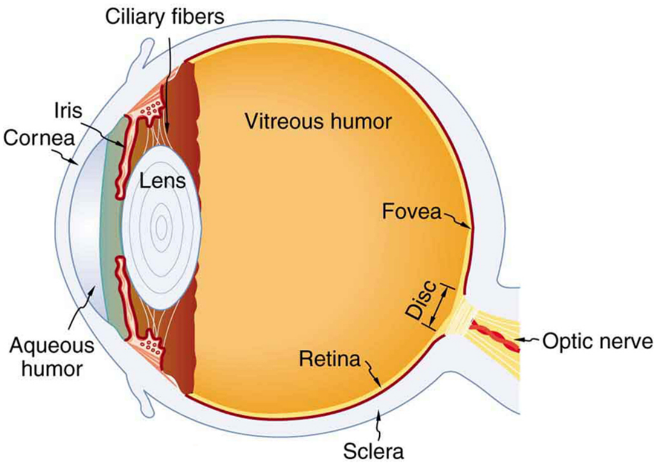
Seeing faces and objects we love and cherish is a delight—one's favorite teddy bear, a picture on the wall, or the sun rising over the mountains. Intricate images help us understand nature and are invaluable for developing techniques and technologies in order to improve the quality of life. The image of a red blood cell that almost fills the cross-sectional area of a tiny capillary makes us wonder how blood makes it through and not get stuck. We are able to see bacteria and viruses and understand their structure. It is the knowledge of physics that provides fundamental understanding and models required to develop new techniques and instruments. Therefore, physics is called an *enabling science*—a science that enables development and advancement in other areas. It is through optics and imaging that physics enables advancement in major areas of biosciences. This chapter illustrates the enabling nature of physics through an understanding of how a human eye is able to see and how we are able to use optical instruments to see beyond what is possible with the naked eye. It is convenient to categorize these instruments on the basis of geometric optics (see Geometric Optics) and wave optics (see Wave Optics).

Physics of the Eye

- Explain the image formation by the eye.
- Explain why peripheral images lack detail and color.
- Define refractive indices.
- Analyze the accommodation of the eye for distant and near vision.

The eye is perhaps the most interesting of all optical instruments. The eye is remarkable in how it forms images and in the richness of detail and color it can detect. However, our eyes commonly need some correction, to reach what is called "normal" vision, but should be called ideal rather than normal. Image formation by our eyes and common vision correction are easy to analyze with the optics discussed in Geometric Optics.

[link] shows the basic anatomy of the eye. The cornea and lens form a system that, to a good approximation, acts as a single thin lens. For clear vision, a real image must be projected onto the light-sensitive retina, which lies at a fixed distance from the lens. The lens of the eye adjusts its power to produce an image on the retina for objects at different distances. The center of the image falls on the fovea, which has the greatest density of light receptors and the greatest acuity (sharpness) in the visual field. The variable opening (or pupil) of the eye along with chemical adaptation allows the eye to detect light intensities from the lowest observable to 10^{10} times greater (without damage). This is an incredible range of detection. Our eyes perform a vast number of functions, such as sense direction, movement, sophisticated colors, and distance. Processing of visual nerve impulses begins with interconnections in the retina and continues in the brain. The optic nerve conveys signals received by the eye to the brain.

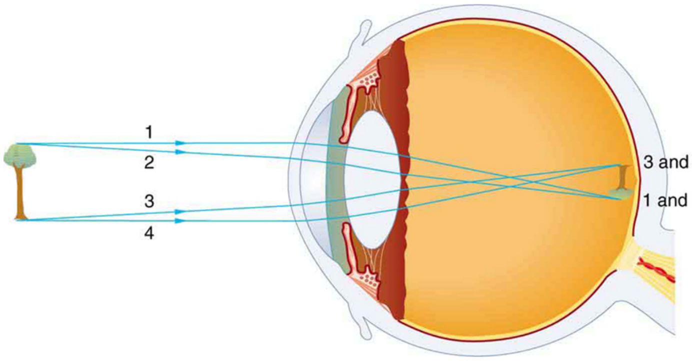


The cornea and lens of an eye act together to form a real image on the light-sensing retina, which has its densest concentration of receptors in the fovea and a blind spot over the optic nerve. The power of the lens of an eye is adjustable to provide an image on the retina for varying object distances. Layers of tissues with varying indices of refraction in the lens are shown here. However, they have been omitted from other pictures for clarity.

Refractive indices are crucial to image formation using lenses. [link] shows refractive indices relevant to the eye. The biggest change in the refractive index, and bending of rays, occurs at the cornea rather than the lens. The ray diagram in [link] shows image formation by the cornea and lens of the eye. The rays bend according to the refractive indices provided in [link]. The cornea provides about two-thirds of the power of the eye, owing to the fact that speed of light changes considerably while traveling from air into cornea. The lens provides the remaining power needed to produce an image on the retina. The cornea and lens can be treated as a single thin lens, even though the light rays pass through several layers of material (such as cornea, aqueous humor, several layers in the lens, and vitreous humor), changing direction at each interface. The image formed is much like the one produced by a single convex lens. This is a case 1 image. Images formed in the eye are inverted but the brain inverts them once more to make them seem upright.

Material	**Index of Refraction**
Water	1.33
Air	1.0
Cornea	1.38
Aqueous humor	1.34
Lens	1.41 average (varies throughout the lens, greatest in center)
Vitreous humor	1.34

Refractive Indices Relevant to the Eye

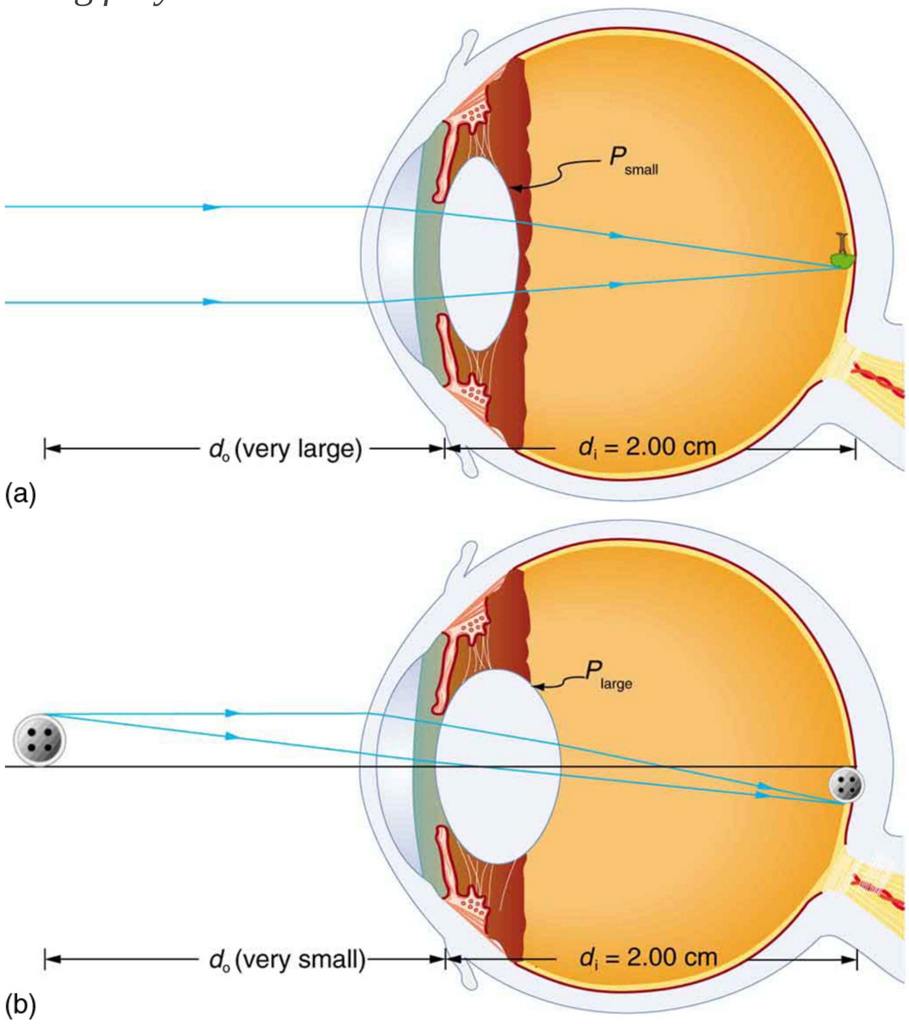


An image is formed on the retina with light rays converging most at the cornea and upon entering and exiting the lens. Rays from the top and bottom of the object are traced and produce an inverted real image on the retina. The distance to the object is drawn smaller than scale.

As noted, the image must fall precisely on the retina to produce clear vision — that is, the image distance d_i must equal the lens-to-retina distance. Because the lens-to-retina distance does not change, the image distance d_i must be the same for objects at all distances. The eye manages this by varying the power (and focal length) of the lens to accommodate for objects at various distances. The process of adjusting the eye's focal length is called **accommodation**. A person with normal (ideal) vision can see objects clearly at distances ranging from 25 cm to essentially infinity. However, although the near point (the shortest distance at which a sharp focus can be obtained) increases with age (becoming meters for some older people), we will consider it to be 25 cm in our treatment here.

[link] shows the accommodation of the eye for distant and near vision. Since light rays from a nearby object can diverge and still enter the eye, the lens must be more converging (more powerful) for close vision than for distant vision. To be more converging, the lens is made thicker by the action of the ciliary muscle surrounding it. The eye is most relaxed when viewing

distant objects, one reason that microscopes and telescopes are designed to produce distant images. Vision of very distant objects is called *totally relaxed*, while close vision is termed *accommodated*, with the closest vision being *fully accommodated*.



Relaxed and accommodated vision for distant and close objects. (a) Light rays from the same point on a distant object must be nearly parallel while entering the eye and more easily converge to produce an image on the retina. (b) Light rays from a nearby object can diverge more and still enter the eye. A more powerful lens is needed to converge them on the retina than if they were parallel.

We will use the thin lens equations to examine image formation by the eye quantitatively. First, note the power of a lens is given as $p = 1/f$, so we rewrite the thin lens equations as

Equation:

$$P = \frac{1}{d_\text{o}} + \frac{1}{d_\text{i}}$$

and

Equation:

$$\frac{h_\text{i}}{h_\text{o}} = -\frac{d_\text{i}}{d_\text{o}} = m.$$

We understand that d_i must equal the lens-to-retina distance to obtain clear vision, and that normal vision is possible for objects at distances $d_\text{o} = 25\text{ cm}$ to infinity.

Note:

Take-Home Experiment: The Pupil

Look at the central transparent area of someone's eye, the pupil, in normal room light. Estimate the diameter of the pupil. Now turn off the lights and darken the room. After a few minutes turn on the lights and promptly estimate the diameter of the pupil. What happens to the pupil as the eye adjusts to the room light? Explain your observations.

The eye can detect an impressive amount of detail, considering how small the image is on the retina. To get some idea of how small the image can be, consider the following example.

Example:
Size of Image on Retina

What is the size of the image on the retina of a 1.20×10^{-2} cm diameter human hair, held at arm's length (60.0 cm) away? Take the lens-to-retina distance to be 2.00 cm.

Strategy

We want to find the height of the image h_i, given the height of the object is $h_\mathrm{o} = 1.20 \times 10^{-2}$ cm. We also know that the object is 60.0 cm away, so that $d_\mathrm{o} = 60.0$ cm. For clear vision, the image distance must equal the lens-to-retina distance, and so $d_\mathrm{i} = 2.00\text{ cm}$. The equation $\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = m$ can be used to find h_i with the known information.

Solution

The only unknown variable in the equation $\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}} = m$ is h_i:

Equation:

$$\frac{h_\mathrm{i}}{h_\mathrm{o}} = -\frac{d_\mathrm{i}}{d_\mathrm{o}}.$$

Rearranging to isolate h_i yields

Equation:

$$h_\mathrm{i} = -h_\mathrm{o} \cdot \frac{d_\mathrm{i}}{d_\mathrm{o}}.$$

Substituting the known values gives

Equation:

$$\begin{aligned} h_\mathrm{i} &= -(1.20 \times 10^{-2}\text{ cm})\frac{2.00\text{ cm}}{60.0\text{ cm}} \\ &= -4.00 \times 10^{-4}\text{ cm}. \end{aligned}$$

Discussion

This truly small image is not the smallest discernible—that is, the limit to visual acuity is even smaller than this. Limitations on visual acuity have to do with the wave properties of light and will be discussed in the next chapter. Some limitation is also due to the inherent anatomy of the eye and processing that occurs in our brain.

Example:

Power Range of the Eye

Calculate the power of the eye when viewing objects at the greatest and smallest distances possible with normal vision, assuming a lens-to-retina distance of 2.00 cm (a typical value).

Strategy

For clear vision, the image must be on the retina, and so $d_{\text{i}} = 2.00 \text{ cm}$ here. For distant vision, $d_{\text{o}} \approx \infty$, and for close vision, $d_{\text{o}} = 25.0 \text{ cm}$, as discussed earlier. The equation $P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$ as written just above, can be used directly to solve for P in both cases, since we know d_{i} and d_{o}. Power has units of diopters, where $1 \text{ D} = 1/\text{m}$, and so we should express all distances in meters.

Solution

For distant vision,

Equation:

$$P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{\infty} + \frac{1}{0.0200 \text{ m}}.$$

Since $1/\infty = 0$, this gives

Equation:

$$P = 0 + 50.0/\text{m} = 50.0 \text{ D (distant vision)}.$$

Now, for close vision,

Equation:

$$\begin{aligned} P &= \tfrac{1}{d_{\text{o}}} + \tfrac{1}{d_{\text{i}}} = \tfrac{1}{0.250 \text{ m}} + \tfrac{1}{0.0200 \text{ m}} \\ &= \tfrac{4.00}{\text{m}} + \tfrac{50.0}{\text{m}} = 4.00 \text{ D} + 50.0 \text{ D} \\ &= 54.0 \text{ D (close vision)}. \end{aligned}$$

Discussion

For an eye with this typical 2.00 cm lens-to-retina distance, the power of the eye ranges from 50.0 D (for distant totally relaxed vision) to 54.0 D (for close fully accommodated vision), which is an 8% increase. This increase in power for close vision is consistent with the preceding discussion and the ray tracing in [link]. An 8% ability to accommodate is considered normal but is typical for people who are about 40 years old. Younger people have greater accommodation ability, whereas older people gradually lose the ability to accommodate. When an optometrist identifies accommodation as a problem in elder people, it is most likely due to stiffening of the lens. The lens of the eye changes with age in ways that tend to preserve the ability to see distant objects clearly but do not allow the eye to accommodate for close vision, a condition called **presbyopia** (literally, elder eye). To correct this vision defect, we place a converging, positive power lens in front of the eye, such as found in reading glasses. Commonly available reading glasses are rated by their power in diopters, typically ranging from 1.0 to 3.5 D.

Section Summary

- Image formation by the eye is adequately described by the thin lens equations:
  Equation:

$$P=\frac{1}{d_\text{o}}+\frac{1}{d_\text{i}}\text{ and }\frac{h_\text{i}}{h_\text{o}}=-\frac{d_\text{i}}{d_\text{o}}=m.$$

- The eye produces a real image on the retina by adjusting its focal length and power in a process called accommodation.
- For close vision, the eye is fully accommodated and has its greatest power, whereas for distant vision, it is totally relaxed and has its smallest power.
- The loss of the ability to accommodate with age is called presbyopia, which is corrected by the use of a converging lens to add power for close vision.

Conceptual Questions

Exercise:

Problem:

If the lens of a person’s eye is removed because of cataracts (as has been done since ancient times), why would you expect a spectacle lens of about 16 D to be prescribed?

Exercise:

Problem:

A cataract is cloudiness in the lens of the eye. Is light dispersed or diffused by it?

Exercise:

Problem:

When laser light is shone into a relaxed normal-vision eye to repair a tear by spot-welding the retina to the back of the eye, the rays entering the eye must be parallel. Why?

Exercise:

Problem:

How does the power of a dry contact lens compare with its power when resting on the tear layer of the eye? Explain.

Exercise:

Problem:

Why is your vision so blurry when you open your eyes while swimming under water? How does a face mask enable clear vision?

Problem Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.

Exercise:

Problem:

What is the power of the eye when viewing an object 50.0 cm away?

Solution:

52.0 D

Exercise:

Problem:

Calculate the power of the eye when viewing an object 3.00 m away.

Exercise:

Problem:

(a) The print in many books averages 3.50 mm in height. How high is the image of the print on the retina when the book is held 30.0 cm from the eye?

(b) Compare the size of the print to the sizes of rods and cones in the fovea and discuss the possible details observable in the letters. (The eye-brain system can perform better because of interconnections and higher order image processing.)

Solution:

(a) -0.233 mm

(b) The size of the rods and the cones is smaller than the image height, so we can distinguish letters on a page.

Exercise:

Problem:

Suppose a certain person’s visual acuity is such that he can see objects clearly that form an image 4.00 μm high on his retina. What is the maximum distance at which he can read the 75.0 cm high letters on the side of an airplane?

Exercise:

Problem:

People who do very detailed work close up, such as jewellers, often can see objects clearly at much closer distance than the normal 25 cm.

(a) What is the power of the eyes of a woman who can see an object clearly at a distance of only 8.00 cm?

(b) What is the size of an image of a 1.00 mm object, such as lettering inside a ring, held at this distance?

(c) What would the size of the image be if the object were held at the normal 25.0 cm distance?

Solution:

(a) $+62.5\ \text{D}$

(b) –$0.250\ \text{mm}$

(c) –$0.0800\ \text{mm}$

Glossary

accommodation
the ability of the eye to adjust its focal length is known as accommodation

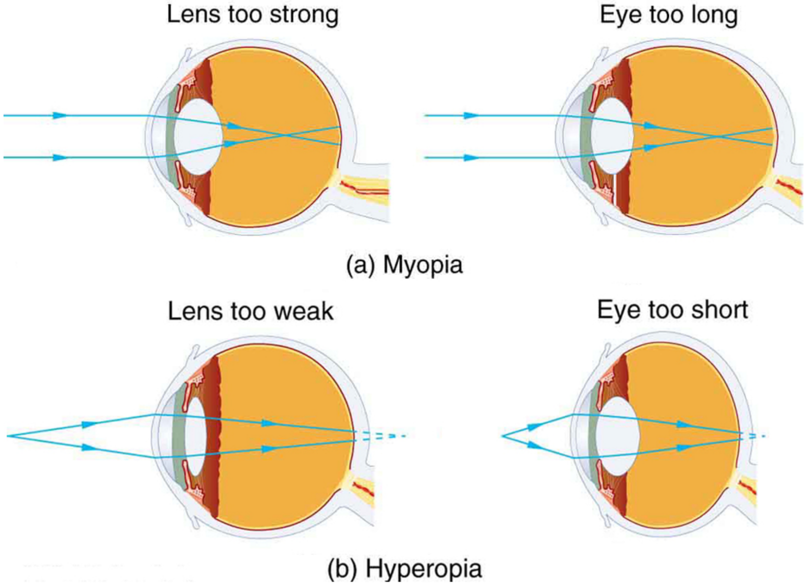
presbyopia

a condition in which the lens of the eye becomes progressively unable to focus on objects close to the viewer

Vision Correction

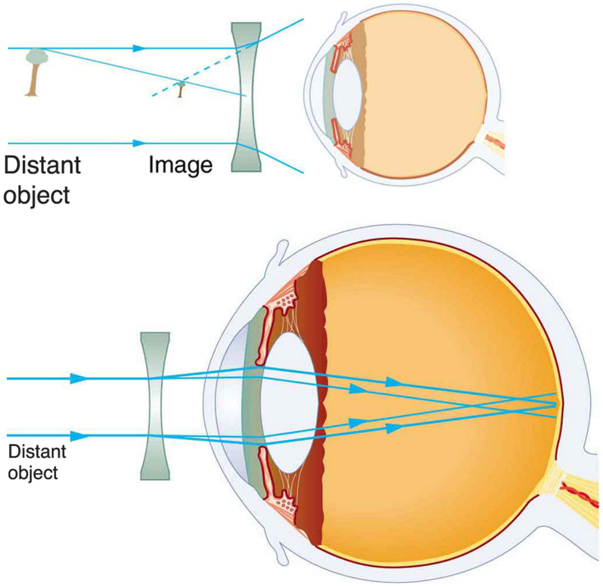
- Identify and discuss common vision defects.
- Explain nearsightedness and farsightedness corrections.
- Explain laser vision correction.

The need for some type of vision correction is very common. Common vision defects are easy to understand, and some are simple to correct. [link] illustrates two common vision defects. **Nearsightedness**, or **myopia**, is the inability to see distant objects clearly while close objects are clear. The eye overconverges the nearly parallel rays from a distant object, and the rays cross in front of the retina. More divergent rays from a close object are converged on the retina for a clear image. The distance to the farthest object that can be seen clearly is called the **far point** of the eye (normally infinity). **Farsightedness**, or **hyperopia**, is the inability to see close objects clearly while distant objects may be clear. A farsighted eye does not converge sufficient rays from a close object to make the rays meet on the retina. Less diverging rays from a distant object can be converged for a clear image. The distance to the closest object that can be seen clearly is called the **near point** of the eye (normally 25 cm).



(a) The nearsighted (myopic) eye converges rays from a distant object in front of the retina; thus, they are diverging when they strike the retina, producing a blurry image. This can be caused by the lens of the eye being too powerful or the length of the eye being too great. (b) The farsighted (hyperopic) eye is unable to converge the rays from a close object by the time they strike the retina, producing blurry close vision. This can be caused by insufficient power in the lens or by the eye being too short.

Since the nearsighted eye over converges light rays, the correction for nearsightedness is to place a diverging spectacle lens in front of the eye. This reduces the power of an eye that is too powerful. Another way of thinking about this is that a diverging spectacle lens produces a case 3 image, which is closer to the eye than the object (see [link]). To determine the spectacle power needed for correction, you must know the person's far point—that is, you must know the greatest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or closer for the nearsighted person to be able to see it clearly. It is worth noting that wearing glasses does not change the eye in any way. The eyeglass lens is simply used to create an image of the object at a distance where the nearsighted person can see it clearly. Whereas someone not wearing glasses can see clearly *objects* that fall between their near point and their far point, someone wearing glasses can see *images* that fall between their near point and their far point.



Correction of nearsightedness requires a diverging lens that compensates for the overconvergence by the eye. The diverging lens produces an image closer to the eye than the object, so that the nearsighted person can see it clearly.

Example:
Correcting Nearsightedness

What power of spectacle lens is needed to correct the vision of a nearsighted person whose far point is 30.0 cm? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

You want this nearsighted person to be able to see very distant objects clearly. That means the spectacle lens must produce an image 30.0 cm from the eye for an object very far away. An image 30.0 cm from the eye will be 28.5 cm to the left of the spectacle lens (see [link]). Therefore, we must get $d_{\text{i}} = -28.5\text{ cm}$ when $d_{\text{o}} \approx \infty$. The image distance is negative, because it is on the same side of the spectacle as the object.

Solution

Since d_{i} and d_{o} are known, the power of the spectacle lens can be found using $P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$ as written earlier:

Equation:

$$P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}} = \frac{1}{\infty} + \frac{1}{-0.285\text{ m}}.$$

Since $1/\infty = 0$, we obtain:

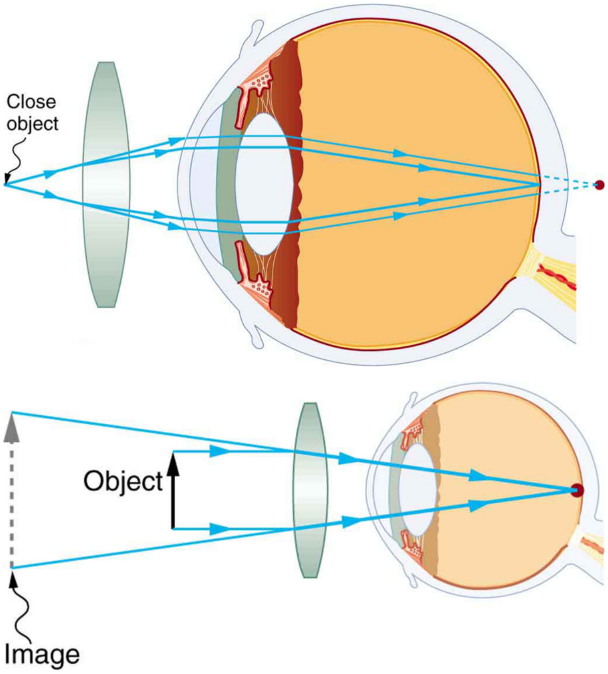
Equation:

$$P = 0 - 3.51/\text{m} = -3.51\text{ D}.$$

Discussion

The negative power indicates a diverging (or concave) lens, as expected. The spectacle produces a case 3 image closer to the eye, where the person can see it. If you examine eyeglasses for nearsighted people, you will find the lenses are thinnest in the center. Additionally, if you examine a prescription for eyeglasses for nearsighted people, you will find that the prescribed power is negative and given in units of diopters.

Since the farsighted eye under converges light rays, the correction for farsightedness is to place a converging spectacle lens in front of the eye. This increases the power of an eye that is too weak. Another way of thinking about this is that a converging spectacle lens produces a case 2 image, which is farther from the eye than the object (see [link]). To determine the spectacle power needed for correction, you must know the person’s near point—that is, you must know the smallest distance at which the person can see clearly. Then the image produced by a spectacle lens must be at this distance or farther for the farsighted person to be able to see it clearly.



Correction of farsightedness uses a converging lens that compensates for the under convergence by the eye. The converging lens produces an image farther from the eye than the object, so that the farsighted person can see it clearly.

Example:
Correcting Farsightedness

What power of spectacle lens is needed to allow a farsighted person, whose near point is 1.00 m, to see an object clearly that is 25.0 cm away? Assume the spectacle (corrective) lens is held 1.50 cm away from the eye by eyeglass frames.

Strategy

When an object is held 25.0 cm from the person’s eyes, the spectacle lens must produce an image 1.00 m away (the near point). An image 1.00 m from the eye will be 98.5 cm to the left of the spectacle lens because the spectacle lens is 1.50 cm from the eye (see [link]). Therefore, $d_{\text{i}} = -98.5\text{ cm}$. The image distance is negative, because it is on the same side of the spectacle as the object. The object is 23.5 cm to the left of the spectacle, so that $d_{\text{o}} = 23.5\text{ cm}$.

Solution

Since d_{i} and d_{o} are known, the power of the spectacle lens can be found using $P = \frac{1}{d_{\text{o}}} + \frac{1}{d_{\text{i}}}$:

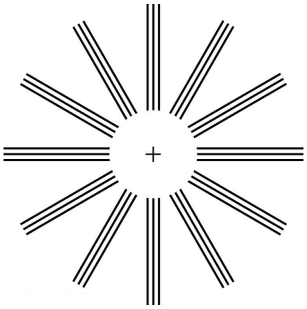
Equation:

$$\begin{array}{lll} P & = & \frac{1}{d_o} + \frac{1}{d_{\text{i}}} = \frac{1}{0.235\text{ m}} + \frac{1}{-0.985\text{ m}} \\ & = & 4.26\text{ D} - 1.02\text{ D} = 3.24\text{ D}. \end{array}$$

Discussion

The positive power indicates a converging (convex) lens, as expected. The convex spectacle produces a case 2 image farther from the eye, where the person can see it. If you examine eyeglasses of farsighted people, you will find the lenses to be thickest in the center. In addition, a prescription of eyeglasses for farsighted people has a prescribed power that is positive.

Another common vision defect is **astigmatism**, an unevenness or asymmetry in the focus of the eye. For example, rays passing through a vertical region of the eye may focus closer than rays passing through a horizontal region, resulting in the image appearing elongated. This is mostly due to irregularities in the shape of the cornea but can also be due to lens irregularities or unevenness in the retina. Because of these irregularities, different parts of the lens system produce images at different locations. The eye-brain system can compensate for some of these irregularities, but they generally manifest themselves as less distinct vision or sharper images along certain axes. [link] shows a chart used to detect astigmatism. Astigmatism can be at least partially corrected with a spectacle having the opposite irregularity of the eye. If an eyeglass prescription has a cylindrical correction, it is there to correct astigmatism. The normal corrections for short- or farsightedness are spherical corrections, uniform along all axes.



This chart can detect astigmatism, unevenness in the focus of the eye. Check each of your eyes separately by looking at the center cross (without spectacles if you wear them). If lines along some axes appear darker or clearer than others, you have an astigmatism.

Contact lenses have advantages over glasses beyond their cosmetic aspects. One problem with glasses is that as the eye moves, it is not at a fixed distance from the spectacle lens. Contacts rest on and move with the eye, eliminating this problem. Because contacts cover a significant portion of the cornea, they provide superior peripheral vision compared with eyeglasses. Contacts also correct some corneal astigmatism caused by surface irregularities. The tear layer between the smooth contact and the cornea fills in the irregularities. Since the index of refraction of the tear layer and the cornea are very similar, you now have a regular optical surface in place of an irregular one. If the curvature of a contact lens is not the same as the cornea (as may be necessary with some individuals to obtain a comfortable fit), the tear layer between the contact and cornea acts as a lens. If the tear layer is thinner in the center than at the edges, it has a negative power, for example. Skilled optometrists will adjust the power of the contact to compensate.

Laser vision correction has progressed rapidly in the last few years. It is the latest and by far the most successful in a series of procedures that correct vision by reshaping the cornea. As noted at the beginning of this section, the cornea accounts for about two-thirds of the power of the eye. Thus, small adjustments of its curvature have the same effect as putting a lens in front of the eye. To a reasonable approximation, the power of multiple lenses placed close together equals the sum of their powers. For example, a concave spectacle lens (for nearsightedness) having $P = -3.00\text{ D}$ has the same effect on vision as reducing the power of the eye itself by 3.00 D. So to correct the eye for nearsightedness, the cornea is flattened to reduce its power. Similarly, to correct for farsightedness, the curvature of the cornea is enhanced to increase the power of the eye—the same effect as the positive power spectacle lens used for farsightedness. Laser vision correction uses high intensity electromagnetic radiation to ablate (to remove material from the surface) and reshape the corneal surfaces.

Today, the most commonly used laser vision correction procedure is *Laser in situ Keratomileusis (LASIK)*. The top layer of the cornea is surgically peeled back and the underlying tissue ablated by multiple bursts of finely controlled ultraviolet radiation produced by an excimer laser. Lasers are used because they not only produce well-focused intense light, but they also emit very pure wavelength electromagnetic radiation that can be controlled more accurately than mixed wavelength light. The 193 nm wavelength UV commonly used is extremely and strongly absorbed by corneal tissue, allowing precise evaporation of very thin layers. A computer controlled program applies more bursts, usually at a rate of 10 per second, to the areas that require deeper removal. Typically a spot less than 1 mm in diameter and about $0.3\ \mu m$ in thickness is removed by each burst. Nearsightedness, farsightedness, and astigmatism can be corrected with an accuracy that produces normal distant vision in more than 90% of the patients, in many cases right away. The corneal flap is replaced; healing takes place rapidly and is nearly painless. More than 1 million Americans per year undergo LASIK (see [link]).

Laser vision correction is being performed using the LASIK procedure. Reshaping of the cornea by laser ablation is based on a careful assessment of the patient’s vision and is computer controlled. The upper corneal layer is temporarily peeled back and minimally disturbed in LASIK, providing for more rapid and less painful healing of the less sensitive tissues below. (credit: U.S. Navy photo by Mass Communication Specialist 1st Class Brien Aho)

Section Summary

- Nearsightedness, or myopia, is the inability to see distant objects and is corrected with a diverging lens to reduce power.
- Farsightedness, or hyperopia, is the inability to see close objects and is corrected with a converging lens to increase power.
- In myopia and hyperopia, the corrective lenses produce images at a distance that the person can see clearly—the far point and near point, respectively.

Conceptual Questions

Exercise:

Problem:

It has become common to replace the cataract-clouded lens of the eye with an internal lens. This intraocular lens can be chosen so that the person has perfect distant vision. Will the person be able to read without glasses? If the person was nearsighted, is the power of the intraocular lens greater or less than the removed lens?

Exercise:

Problem:

If the cornea is to be reshaped (this can be done surgically or with contact lenses) to correct myopia, should its curvature be made greater or smaller? Explain. Also explain how hyperopia can be corrected.

Exercise:

Problem:

If there is a fixed percent uncertainty in LASIK reshaping of the cornea, why would you expect those people with the greatest correction to have a poorer chance of normal distant vision after the procedure?

Exercise:

Problem:

A person with presbyopia has lost some or all of the ability to accommodate the power of the eye. If such a person’s distant vision is corrected with LASIK, will she still need reading glasses? Explain.

Problem Exercises

Exercise:

Problem:

What is the far point of a person whose eyes have a relaxed power of 50.5 D?

Solution:

2.00 m

Exercise:

Problem:

What is the near point of a person whose eyes have an accommodated power of 53.5 D?

Exercise:

Problem:

(a) A laser vision correction reshaping the cornea of a myopic patient reduces the power of his eye by 9.00 D, with a $\pm 5.0\%$ uncertainty in the final correction. What is the range of diopters for spectacle lenses that this person might need after LASIK procedure? (b) Was the person nearsighted or farsighted before the procedure? How do you know?

Solution:

(a) ± 0.45 D

(b) The person was nearsighted because the patient was myopic and the power was reduced.

Exercise:

Problem:

In a LASIK vision correction, the power of a patient's eye is increased by 3.00 D. Assuming this produces normal close vision, what was the patient's near point before the procedure?

Exercise:

Problem:

What was the previous far point of a patient who had laser vision correction that reduced the power of her eye by 7.00 D, producing normal distant vision for her?

Solution:

0.143 m

Exercise:

Problem:

A severely myopic patient has a far point of 5.00 cm. By how many diopters should the power of his eye be reduced in laser vision correction to obtain normal distant vision for him?

Exercise:

Problem:

A student’s eyes, while reading the blackboard, have a power of 51.0 D. How far is the board from his eyes?

Solution:

1.00 m

Exercise:

Problem:

The power of a physician’s eyes is 53.0 D while examining a patient. How far from her eyes is the feature being examined?

Exercise:

Problem:

A young woman with normal distant vision has a 10.0% ability to accommodate (that is, increase) the power of her eyes. What is the closest object she can see clearly?

Solution:

20.0 cm

Exercise:

Problem:

The far point of a myopic administrator is 50.0 cm. (a) What is the relaxed power of his eyes? (b) If he has the normal 8.00% ability to accommodate, what is the closest object he can see clearly?

Exercise:

Problem:

A very myopic man has a far point of 20.0 cm. What power contact lens (when on the eye) will correct his distant vision?

Solution:

–5.00 D

Exercise:

Problem:

Repeat the previous problem for eyeglasses held 1.50 cm from the eyes.

Exercise:

Problem:

A myopic person sees that her contact lens prescription is –4.00 D. What is her far point?

Solution:

25.0 cm

Exercise:

Problem:

Repeat the previous problem for glasses that are 1.75 cm from the eyes.

Exercise:

Problem:

The contact lens prescription for a mildly farsighted person is 0.750 D, and the person has a near point of 29.0 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

Solution:

–0.198 D

Exercise:

Problem:

A nearsighted man cannot see objects clearly beyond 20 cm from his eyes. How close must he stand to a mirror in order to see what he is doing when he shaves?

Exercise:

Problem:

A mother sees that her child’s contact lens prescription is 0.750 D. What is the child’s near point?

Solution:

30.8 cm

Exercise:

Problem:

Repeat the previous problem for glasses that are 2.20 cm from the eyes.

Exercise:

Problem:

The contact lens prescription for a nearsighted person is –4.00 D and the person has a far point of 22.5 cm. What is the power of the tear layer between the cornea and the lens if the correction is ideal, taking the tear layer into account?

Solution:

–0.444 D

Exercise:

Problem: **Unreasonable Results**

A boy has a near point of 50 cm and a far point of 500 cm. Will a –4.00 D lens correct his far point to infinity?

Glossary

nearsightedness
: another term for myopia, a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

myopia
: a visual defect in which distant objects appear blurred because their images are focused in front of the retina rather than being focused on the retina

far point

the object point imaged by the eye onto the retina in an unaccommodated eye

farsightedness

another term for hyperopia, the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

hyperopia

the condition of an eye where incoming rays of light reach the retina before they converge into a focused image

near point

the point nearest the eye at which an object is accurately focused on the retina at full accommodation

astigmatism

the result of an inability of the cornea to properly focus an image onto the retina

laser vision correction

a medical procedure used to correct astigmatism and eyesight deficiencies such as myopia and hyperopia

Color and Color Vision

- Explain the simple theory of color vision.
- Outline the coloring properties of light sources.
- Describe the retinex theory of color vision.

The gift of vision is made richer by the existence of color. Objects and lights abound with thousands of hues that stimulate our eyes, brains, and emotions. Two basic questions are addressed in this brief treatment—what does color mean in scientific terms, and how do we, as humans, perceive it?

Simple Theory of Color Vision

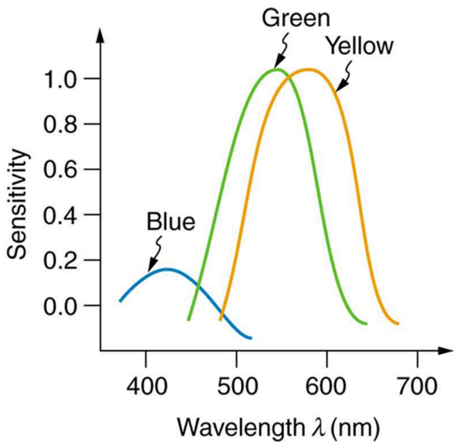
We have already noted that color is associated with the wavelength of visible electromagnetic radiation. When our eyes receive pure-wavelength light, we tend to see only a few colors. Six of these (most often listed) are red, orange, yellow, green, blue, and violet. These are the rainbow of colors produced when white light is dispersed according to different wavelengths. There are thousands of other **hues** that we can perceive. These include brown, teal, gold, pink, and white. One simple theory of color vision implies that all these hues are our eye’s response to different combinations of wavelengths. This is true to an extent, but we find that color perception is even subtler than our eye’s response for various wavelengths of light.

The two major types of light-sensing cells (photoreceptors) in the retina are **rods and cones**. Rods are more sensitive than cones by a factor of about 1000 and are solely responsible for peripheral vision as well as vision in very dark environments. They are also important for motion detection. There are about 120 million rods in the human retina. Rods do not yield color information. You may notice that you lose color vision when it is very dark, but you retain the ability to discern grey scales.

Note:
Take-Home Experiment: Rods and Cones

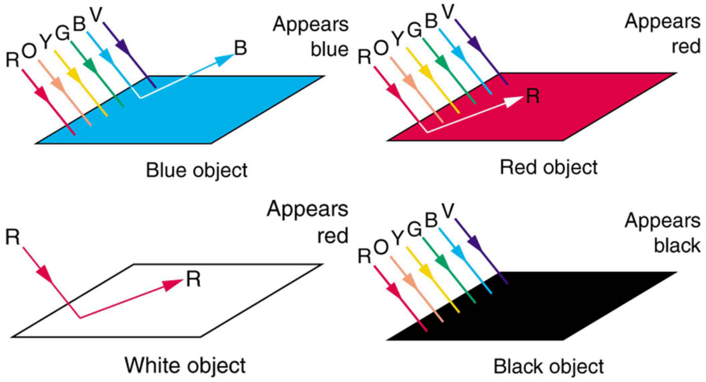
1. Go into a darkened room from a brightly lit room, or from outside in the Sun. How long did it take to start seeing shapes more clearly? What about color? Return to the bright room. Did it take a few minutes before you could see things clearly?
2. Demonstrate the sensitivity of foveal vision. Look at the letter G in the word ROGERS. What about the clarity of the letters on either side of G?

Cones are most concentrated in the fovea, the central region of the retina. There are no rods here. The fovea is at the center of the macula, a 5 mm diameter region responsible for our central vision. The cones work best in bright light and are responsible for high resolution vision. There are about 6 million cones in the human retina. There are three types of cones, and each type is sensitive to different ranges of wavelengths, as illustrated in [link]. A **simplified theory of color vision** is that there are three *primary colors* corresponding to the three types of cones. The thousands of other hues that we can distinguish among are created by various combinations of stimulations of the three types of cones. Color television uses a three-color system in which the screen is covered with equal numbers of red, green, and blue phosphor dots. The broad range of hues a viewer sees is produced by various combinations of these three colors. For example, you will perceive yellow when red and green are illuminated with the correct ratio of intensities. White may be sensed when all three are illuminated. Then, it would seem that all hues can be produced by adding three primary colors in various proportions. But there is an indication that color vision is more sophisticated. There is no unique set of three primary colors. Another set that works is yellow, green, and blue. A further indication of the need for a more complex theory of color vision is that various different combinations can produce the same hue. Yellow can be sensed with yellow light, or with a combination of red and green, and also with white light from which violet has been removed. The three-primary-colors aspect of color vision is well established; more sophisticated theories expand on it rather than deny it.



The image shows the relative sensitivity of the three types of cones, which are named according to wavelengths of greatest sensitivity. Rods are about 1000 times more sensitive, and their curve peaks at about 500 nm. Evidence for the three types of cones comes from direct measurements in animal and human eyes and testing of color blind people.

Consider why various objects display color—that is, why are feathers blue and red in a crimson rosella? The *true color of an object* is defined by its absorptive or reflective characteristics. [link] shows white light falling on three different objects, one pure blue, one pure red, and one black, as well as pure red light falling on a white object. Other hues are created by more complex absorption characteristics. Pink, for example on a galah cockatoo, can be due to weak absorption of all colors except red. An object can appear a different color under non-white illumination. For example, a pure blue object illuminated with pure red light will *appear* black, because it absorbs all the red light falling on it. But, the true color of the object is blue, which is independent of illumination.



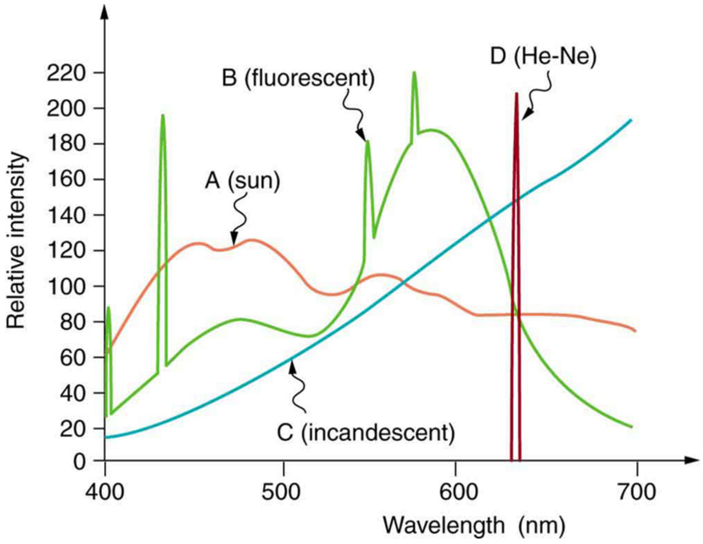
Absorption characteristics determine the true color of an object. Here, three objects are illuminated by white light, and one by pure red light. White is the equal mixture of all visible wavelengths; black is the absence of light.

Similarly, *light sources have colors* that are defined by the wavelengths they produce. A helium-neon laser emits pure red light. In fact, the phrase “pure red light” is defined by having a sharp constrained spectrum, a characteristic of laser light. The Sun produces a broad yellowish spectrum, fluorescent lights emit bluish-white light, and incandescent lights emit reddish-white hues as seen in [link]. As you would expect, you sense these colors when viewing the light source directly or when illuminating a white object with them. All of this fits neatly into the simplified theory that a combination of wavelengths produces various hues.

Note:

Take-Home Experiment: Exploring Color Addition

This activity is best done with plastic sheets of different colors as they allow more light to pass through to our eyes. However, thin sheets of paper and fabric can also be used. Overlay different colors of the material and hold them up to a white light. Using the theory described above, explain the colors you observe. You could also try mixing different crayon colors.

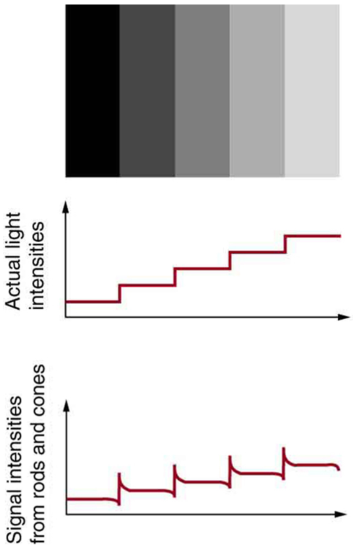


Emission spectra for various light sources are shown. Curve A is average sunlight at Earth’s surface, curve B is light from a fluorescent lamp, and curve C is the output of an incandescent light. The spike for a helium-neon laser (curve D) is due to its pure wavelength emission. The spikes in the fluorescent output are due to atomic spectra—a topic that will be explored later.

Color Constancy and a Modified Theory of Color Vision

The eye-brain color-sensing system can, by comparing various objects in its view, perceive the true color of an object under varying lighting conditions—an ability that is called **color constancy**. We can sense that a white tablecloth, for example, is white whether it is illuminated by sunlight, fluorescent light, or candlelight. The wavelengths entering the eye are quite different in each case, as the graphs in [link] imply, but our color vision can detect the true color by comparing the tablecloth with its surroundings.

Theories that take color constancy into account are based on a large body of anatomical evidence as well as perceptual studies. There are nerve connections among the light receptors on the retina, and there are far fewer nerve connections to the brain than there are rods and cones. This means that there is signal processing in the eye before information is sent to the brain. For example, the eye makes comparisons between adjacent light receptors and is very sensitive to edges as seen in [link]. Rather than responding simply to the light entering the eye, which is uniform in the various rectangles in this figure, the eye responds to the edges and senses false darkness variations.



The importance of edges is

shown. Although the grey strips are uniformly shaded, as indicated by the graph immediately below them, they do not appear uniform at all. Instead, they are perceived darker on the dark side and lighter on the light side of the edge, as shown in the bottom graph. This is due to nerve impulse processing in the eye.

One theory that takes various factors into account was advanced by Edwin Land (1909 – 1991), the creative founder of the Polaroid Corporation. Land proposed, based partly on his many elegant experiments, that the three types of cones are organized into systems called **retinexes**. Each retinex forms an image that is compared with the others, and the eye-brain system thus can compare a candle-illuminated white table cloth with its generally reddish surroundings and determine that it is actually white. This **retinex theory of color vision** is an example of modified theories of color vision that attempt to account for its subtleties. One striking experiment performed by Land demonstrates that some type of image comparison may produce color vision. Two pictures are taken of a scene on black-and-white film, one using a red filter, the other a blue filter. Resulting black-and-white slides are then projected and superimposed on a screen, producing a black-and-white image, as expected. Then a red filter is placed in front of the slide taken with a red filter, and the images are again superimposed on a screen. You would expect an image in various shades of pink, but instead, the image appears to humans in full color with all the hues of the original scene. This implies that color vision can be induced by comparison of the black-and-white and red images. Color vision is not completely understood or explained, and the retinex theory is not totally accepted. It is apparent that color vision is much subtler than what a first look might imply.

Note:

PhET Explorations: Color Vision

Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.

https://phet.colorado.edu/sims/html/color-vision/latest/color-vision_en.html

Section Summary

- The eye has four types of light receptors—rods and three types of color-sensitive cones.
- The rods are good for night vision, peripheral vision, and motion changes, while the cones are responsible for central vision and color.
- We perceive many hues, from light having mixtures of wavelengths.
- A simplified theory of color vision states that there are three primary colors, which correspond to the three types of cones, and that various combinations of the primary colors produce all the hues.
- The true color of an object is related to its relative absorption of various wavelengths of light. The color of a light source is related to the wavelengths it produces.

- Color constancy is the ability of the eye-brain system to discern the true color of an object illuminated by various light sources.
- The retinex theory of color vision explains color constancy by postulating the existence of three retinexes or image systems, associated with the three types of cones that are compared to obtain sophisticated information.

Conceptual Questions

Exercise:

Problem:

A pure red object on a black background seems to disappear when illuminated with pure green light. Explain why.

Exercise:

Problem: What is color constancy, and what are its limitations?

Exercise:

Problem:

There are different types of color blindness related to the malfunction of different types of cones. Why would it be particularly useful to study those rare individuals who are color blind only in one eye or who have a different type of color blindness in each eye?

Exercise:

Problem:

Propose a way to study the function of the rods alone, given they can sense light about 1000 times dimmer than the cones.

Glossary

hues

identity of a color as it relates specifically to the spectrum

rods and cones

two types of photoreceptors in the human retina; rods are responsible for vision at low light levels, while cones are active at higher light levels

simplified theory of color vision

a theory that states that there are three primary colors, which correspond to the three types of cones

color constancy

a part of the visual perception system that allows people to perceive color in a variety of conditions and to see some consistency in the color

retinex

a theory proposed to explain color and brightness perception and constancies; is a combination of the words retina and cortex, which are the two areas responsible for the processing of visual information

retinex theory of color vision

the ability to perceive color in an ambient-colored environment

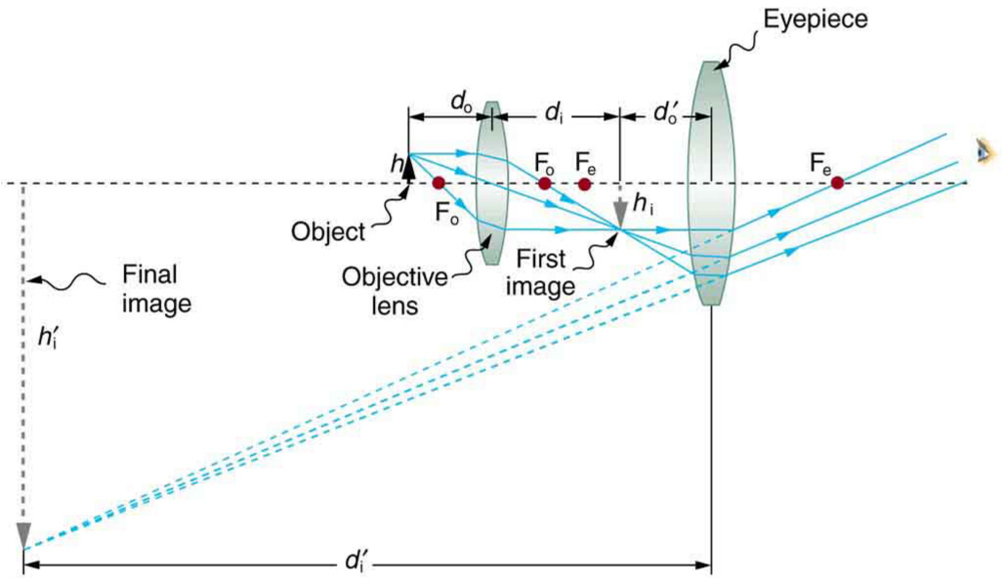
Microscopes

- Investigate different types of microscopes.
- Learn how image is formed in a compound microscope.

Although the eye is marvelous in its ability to see objects large and small, it obviously has limitations to the smallest details it can detect. Human desire to see beyond what is possible with the naked eye led to the use of optical instruments. In this section we will examine microscopes, instruments for enlarging the detail that we cannot see with the unaided eye. The microscope is a multiple-element system having more than a single lens or mirror. (See [link]) A microscope can be made from two convex lenses. The image formed by the first element becomes the object for the second element. The second element forms its own image, which is the object for the third element, and so on. Ray tracing helps to visualize the image formed. If the device is composed of thin lenses and mirrors that obey the thin lens equations, then it is not difficult to describe their behavior numerically.

Multiple lenses and mirrors are used in this microscope. (credit: U.S. Navy photo by Tom Watanabe)

Microscopes were first developed in the early 1600s by eyeglass makers in The Netherlands and Denmark. The simplest **compound microscope** is constructed from two convex lenses as shown schematically in [link]. The first lens is called the **objective lens**, and has typical magnification values from $5\times$ to $100\times$. In standard microscopes, the objectives are mounted such that when you switch between objectives, the sample remains in focus. Objectives arranged in this way are described as parfocal. The second, the **eyepiece**, also referred to as the ocular, has several lenses which slide inside a cylindrical barrel. The focusing ability is provided by the movement of both the objective lens and the eyepiece. The purpose of a microscope is to magnify small objects, and both lenses contribute to the final magnification. Additionally, the final enlarged image is produced in a location far enough from the observer to be easily viewed, since the eye cannot focus on objects or images that are too close.



A compound microscope composed of two lenses, an objective and an eyepiece. The objective forms a case 1 image that is larger than the object. This first image is the object for the eyepiece. The eyepiece forms a case 2 final image that is further magnified.

To see how the microscope in [link] forms an image, we consider its two lenses in succession. The object is slightly farther away from the objective lens than its focal length f_o, producing a case 1 image that is larger than the object. This first image is the object for the second lens, or eyepiece. The eyepiece is intentionally located so it can further magnify the image. The eyepiece is placed so that the first image is closer to it than its focal length f_{e}. Thus the eyepiece acts as a magnifying glass, and the final image is made even larger. The final image remains inverted, but it is farther from the observer, making it easy to view (the eye is most relaxed when viewing distant objects and normally cannot focus closer than 25 cm). Since each lens produces a magnification that multiplies the height of the image, it is apparent that the overall magnification m is the product of the individual magnifications:

Equation:

$$m = m_{\text{o}}m_{\text{e}},$$

where m_{o} is the magnification of the objective and m_{e} is the magnification of the eyepiece. This equation can be generalized for any combination of thin lenses and mirrors that obey the thin lens equations.

Note:

Overall Magnification

The overall magnification of a multiple-element system is the product of the individual magnifications of its elements.

Example:

Microscope Magnification

Calculate the magnification of an object placed 6.20 mm from a compound microscope that has a 6.00 mm focal length objective and a 50.0 mm focal length eyepiece. The objective and eyepiece are separated by 23.0 cm.

Strategy and Concept

This situation is similar to that shown in [link]. To find the overall magnification, we must find the magnification of the objective, then the magnification of the eyepiece. This involves using the thin lens equation.

Solution

The magnification of the objective lens is given as

Equation:

$$m_{\text{o}} = -\frac{d_{\text{i}}}{d_{\text{o}}},$$

where d_{o} and d_{i} are the object and image distances, respectively, for the objective lens as labeled in [link]. The object distance is given to be $d_{\text{o}} = 6.20\ \text{mm}$, but the image distance d_{i} is not known. Isolating d_{i}, we have

Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{f_{\text{o}}} - \frac{1}{d_{\text{o}}},$$

where f_{o} is the focal length of the objective lens. Substituting known values gives

Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{6.00\ \text{mm}} - \frac{1}{6.20\ \text{mm}} = \frac{0.00538}{\text{mm}}.$$

We invert this to find d_{i}:

Equation:

$$d_{\text{i}} = 186\ \text{mm}.$$

Substituting this into the expression for m_{o} gives

Equation:

$$m_{\text{o}} = -\frac{d_{\text{i}}}{d_{\text{o}}} = -\frac{186\ \text{mm}}{6.20\ \text{mm}} = -30.0.$$

Now we must find the magnification of the eyepiece, which is given by

Equation:

$$m_{\text{e}} = -\frac{d_{\text{i}}\prime}{d_{\text{o}}\prime},$$

where $d_{\text{i}}\prime$ and $d_{\text{o}}\prime$ are the image and object distances for the eyepiece (see [link]). The object distance is the distance of the first image from the eyepiece. Since the first image is 186 mm to the right of the objective and the eyepiece is 230 mm to the right of the objective, the object distance is $d_{\text{o}}\prime = 230\text{ mm} - 186\text{ mm} = 44.0\text{ mm}$. This places the first image closer to the eyepiece than its focal length, so that the eyepiece will form a case 2 image as shown in the figure. We still need to find the location of the final image $d_{\text{i}}\prime$ in order to find the magnification. This is done as before to obtain a value for $1/d_{\text{i}}\prime$:

Equation:

$$\frac{1}{d_{\text{i}}\prime} = \frac{1}{f_{\text{e}}} - \frac{1}{d_{\text{o}}\prime} = \frac{1}{50.0\text{ mm}} - \frac{1}{44.0\text{ mm}} = -\frac{0.00273}{\text{mm}}.$$

Inverting gives

Equation:

$$d_{\text{i}}\prime = -\frac{\text{mm}}{0.00273} = -367\text{ mm}.$$

The eyepiece’s magnification is thus

Equation:

$$m_{\text{e}} = -\frac{d_{\text{i}}\prime}{d_{\text{o}}\prime} = -\frac{-367\text{ mm}}{44.0\text{ mm}} = 8.33.$$

So the overall magnification is

Equation:

$$m = m_{\text{o}} m_{\text{e}} = (-30.0)(8.33) = -250.$$

Discussion

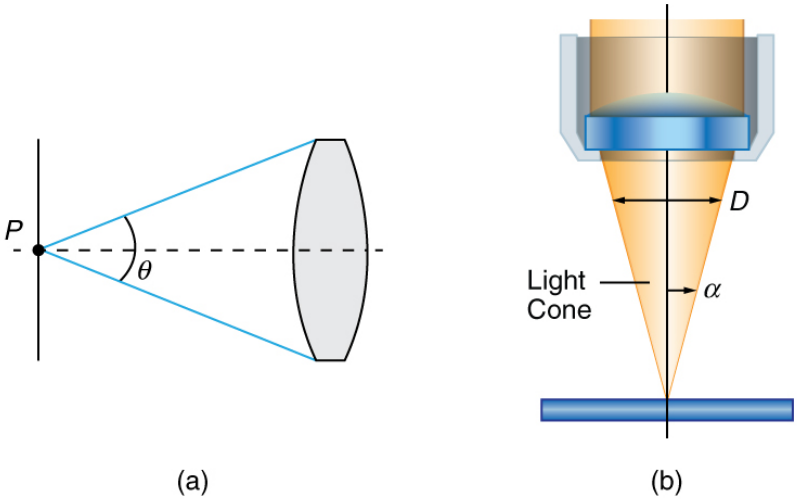
Both the objective and the eyepiece contribute to the overall magnification, which is large and negative, consistent with [link], where the image is seen to be large and inverted. In this case, the image is virtual and inverted, which cannot happen for a single element (case 2 and case 3 images for single elements are virtual and upright). The final image is 367 mm (0.367 m) to the left of the eyepiece. Had the eyepiece been placed farther from the objective, it could have formed a case 1 image to the right. Such an image could be projected on a screen, but it would be behind the head of the person in the figure and not appropriate for direct viewing. The procedure used to solve this example is applicable in any multiple-element system. Each element is treated in turn, with each forming an image that becomes the object for the next element. The process is not more difficult than for single lenses or mirrors, only lengthier.

Normal optical microscopes can magnify up to $1500\times$ with a theoretical resolution of $-0.2\ \mu\text{m}$. The lenses can be quite complicated and are composed of multiple elements to reduce aberrations. Microscope objective lenses are particularly important as they primarily gather light from the specimen. Three parameters describe microscope objectives: the **numerical aperture** (NA), the magnification (m), and the working distance. The NA is related to the light gathering ability of a lens and is obtained using the angle of acceptance θ formed by the maximum cone of rays focusing on the specimen (see [link](a)) and is given by

Equation:

$$\text{NA} = n\ \text{sin}\ \text{α},$$

where n is the refractive index of the medium between the lens and the specimen and $\alpha = \theta/2$. As the angle of acceptance given by θ increases, NA becomes larger and more light is gathered from a smaller focal region giving higher resolution. A 0.75NA objective gives more detail than a $0.10NA$ objective.



(a) The numerical aperture (NA) of a microscope objective lens refers to the light-gathering ability of the lens and is calculated using half the angle of acceptance θ. (b) Here, α is half the acceptance angle for light rays from a specimen entering a camera lens, and D is the diameter of the aperture that controls the light entering the lens.

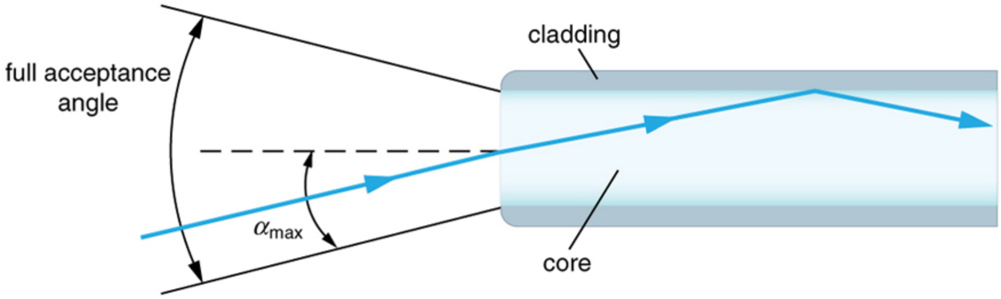
While the numerical aperture can be used to compare resolutions of various objectives, it does not indicate how far the lens could be from the specimen. This is specified by the “working distance,” which is the distance (in mm usually) from the front lens element of the objective to the specimen, or cover glass. The higher the NA the closer the lens will be to the specimen and the more chances there are of breaking the cover slip and damaging both the specimen and the lens. The focal length of an objective lens is different than the working distance. This is because objective lenses are made of a combination of lenses and the focal length is measured from inside the barrel. The working distance is a parameter that microscopists can use more readily as it is measured from the outermost lens. The working distance decreases as the NA and magnification both increase.

The term $f/\#$ in general is called the f-number and is used to denote the light per unit area reaching the image plane. In photography, an image of an object at infinity is formed at the focal point and the f-number is given by the ratio of the focal length f of the lens and the diameter D of the aperture controlling the light into the lens (see [link](b)). If the acceptance angle is small the NA of the lens can also be used as given below.

Equation:

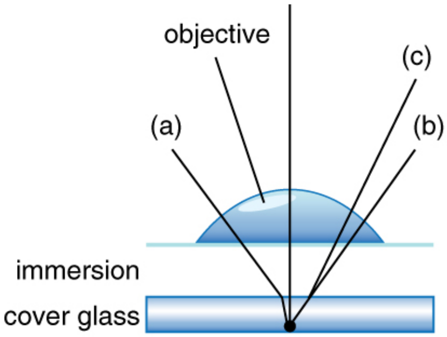
$$f/\# = \frac{f}{D} \approx \frac{1}{2\text{NA}}.$$

As the f-number decreases, the camera is able to gather light from a larger angle, giving wide-angle photography. As usual there is a trade-off. A greater $f/\#$ means less light reaches the image plane. A setting of $f/16$ usually allows one to take pictures in bright sunlight as the aperture diameter is small. In optical fibers, light needs to be focused into the fiber. [link] shows the angle used in calculating the NA of an optical fiber.



Light rays enter an optical fiber. The numerical aperture of the optical fiber can be determined by using the angle α_{max}.

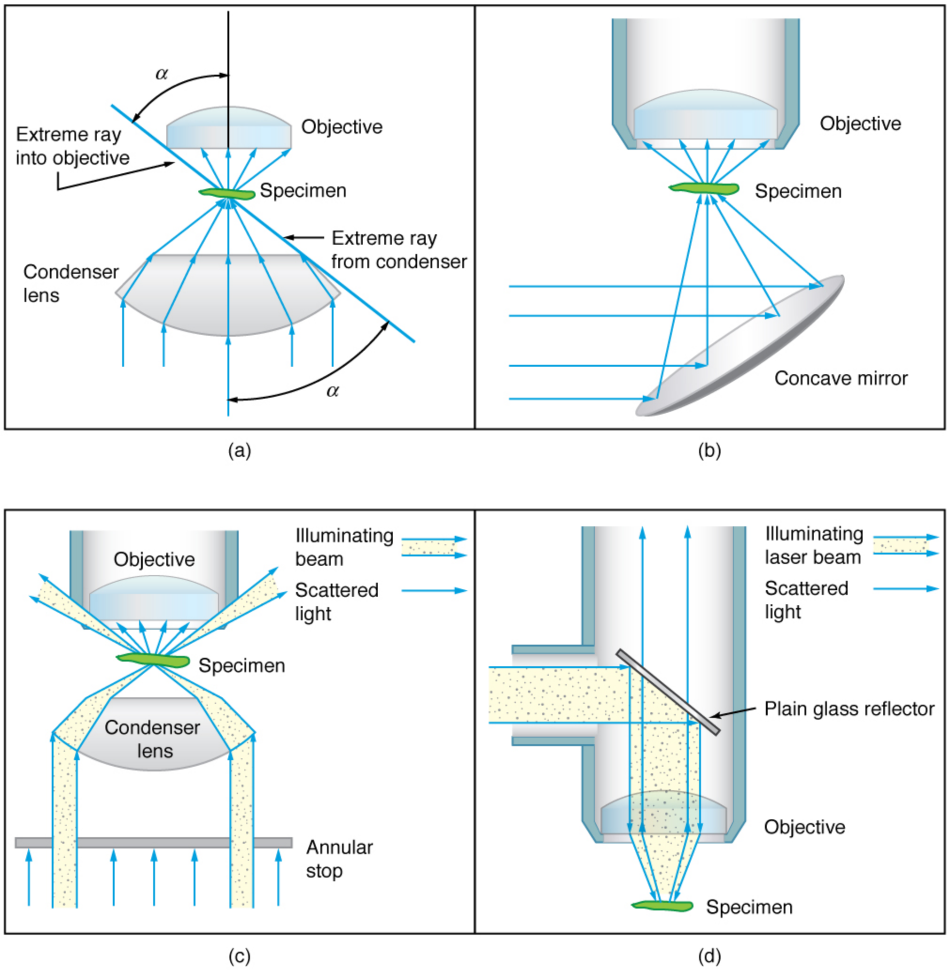
Can the NA be larger than 1.00? The answer is ‘yes’ if we use immersion lenses in which a medium such as oil, glycerine or water is placed between the objective and the microscope cover slip. This minimizes the mismatch in refractive indices as light rays go through different media, generally providing a greater light-gathering ability and an increase in resolution. [link] shows light rays when using air and immersion lenses.



Light rays from a specimen entering the objective. Paths for immersion medium of air (a), water (b) $(n = 1.33)$, and oil (c) $(n = 1.51)$ are shown. The water and oil immersions allow more rays to enter the objective, increasing the resolution.

When using a microscope we do not see the entire extent of the sample. Depending on the eyepiece and objective lens we see a restricted region which we say is the field of view. The objective is then manipulated in two-dimensions above the sample to view other regions of the sample. Electronic scanning of either the objective or the sample is used in scanning microscopy. The image formed at each point during the scanning is combined using a computer to generate an image of a larger region of the sample at a selected magnification.

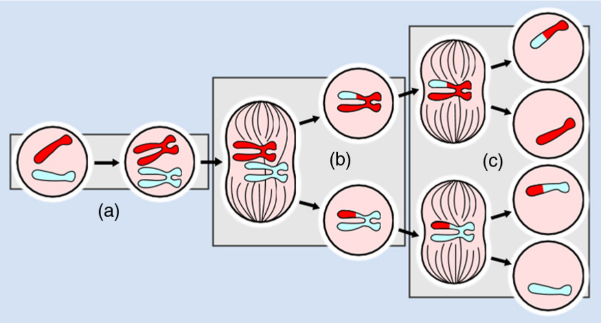
When using a microscope, we rely on gathering light to form an image. Hence most specimens need to be illuminated, particularly at higher magnifications, when observing details that are so small that they reflect only small amounts of light. To make such objects easily visible, the intensity of light falling on them needs to be increased. Special illuminating systems called condensers are used for this purpose. The type of condenser that is suitable for an application depends on how the specimen is examined, whether by transmission, scattering or reflecting. See [link] for an example of each. White light sources are common and lasers are often used. Laser light illumination tends to be quite intense and it is important to ensure that the light does not result in the degradation of the specimen.



Illumination of a specimen in a microscope. (a) Transmitted light from a condenser lens. (b) Transmitted light from a mirror condenser. (c) Dark field illumination by scattering (the illuminating beam misses the objective lens). (d) High magnification illumination with reflected light – normally laser light.

We normally associate microscopes with visible light but x ray and electron microscopes provide greater resolution. The focusing and basic physics is the same as that just described, even though the lenses require different technology. The electron microscope requires vacuum chambers so that the electrons can proceed unheeded. Magnifications of 50 million times provide the ability to determine positions of individual atoms within materials. An electron microscope is shown in [link]. We do not use our eyes to form images; rather images are recorded electronically and displayed on computers. In fact observing and saving images formed by optical microscopes on computers is now done routinely. Video recordings of what occurs in a microscope can be made for viewing by many people at later dates. Physics provides the science and tools needed to generate the sequence of time-lapse images of meiosis similar to the sequence sketched in [link].

An electron microscope has the capability to image individual atoms on a material. The microscope uses vacuum technology, sophisticated detectors and state of the art image processing software. (credit: Dave Pape)



The image shows a sequence of events that takes place during meiosis. (credit: PatríciaR, Wikimedia Commons; National Center for Biotechnology Information)

Note:
Take-Home Experiment: Make a Lens
Look through a clear glass or plastic bottle and describe what you see. Now fill the bottle with water and describe what you see. Use the water bottle as a lens to produce the image of a bright object and estimate the focal length of the water bottle lens. How is the focal length a function of the depth of water in the bottle?

Section Summary

- The microscope is a multiple-element system having more than a single lens or mirror.
- Many optical devices contain more than a single lens or mirror. These are analysed by considering each element sequentially. The image formed by the first is the object for the second, and so on. The same ray tracing and thin lens techniques apply to each lens element.

- The overall magnification of a multiple-element system is the product of the magnifications of its individual elements. For a two-element system with an objective and an eyepiece, this is
  Equation:

  $$m=m_{\text{o}}m_{\text{e}},$$

  where m_{o} is the magnification of the objective and m_{e} is the magnification of the eyepiece, such as for a microscope.
- Microscopes are instruments for allowing us to see detail we would not be able to see with the unaided eye and consist of a range of components.
- The eyepiece and objective contribute to the magnification. The numerical aperture (NA) of an objective is given by
  Equation:

  $$\text{NA}=n\sin\alpha$$

  where n is the refractive index and α the angle of acceptance.
- Immersion techniques are often used to improve the light gathering ability of microscopes. The specimen is illuminated by transmitted, scattered or reflected light though a condenser.
- The $f\ /\#$ describes the light gathering ability of a lens. It is given by
  Equation:

  $$f\ /\#=\frac{f}{D}\approx\frac{1}{2\,NA}.$$

Conceptual Questions

Exercise:

Problem:

Geometric optics describes the interaction of light with macroscopic objects. Why, then, is it correct to use geometric optics to analyse a microscope’s image?

Exercise:

Problem:

The image produced by the microscope in [link] cannot be projected. Could extra lenses or mirrors project it? Explain.

Exercise:

Problem:

Why not have the objective of a microscope form a case 2 image with a large magnification? (Hint: Consider the location of that image and the difficulty that would pose for using the eyepiece as a magnifier.)

Exercise:

Problem: What advantages do oil immersion objectives offer?

Exercise:

Problem:

How does the NA of a microscope compare with the NA of an optical fiber?

Problem Exercises

Exercise:

Problem:

A microscope with an overall magnification of 800 has an objective that magnifies by 200. (a) What is the magnification of the eyepiece? (b) If there are two other objectives that can be used, having magnifications of 100 and 400, what other total magnifications are possible?

Solution:

(a) 4.00

(b) 1600

Exercise:

Problem:

(a) What magnification is produced by a 0.150 cm focal length microscope objective that is 0.155 cm from the object being viewed? (b) What is the overall magnification if an $8\times$ eyepiece (one that produces a magnification of 8.00) is used?

Exercise:

Problem:

(a) Where does an object need to be placed relative to a microscope for its 0.500 cm focal length objective to produce a magnification of –400 ? (b) Where should the 5.00 cm focal length eyepiece be placed to produce a further fourfold (4.00) magnification?

Solution:

(a) 0.501 cm

(b) Eyepiece should be 204 cm behind the objective lens.

Exercise:

Problem:

You switch from a $1.40 NA\ 60\times$ oil immersion objective to a $1.40 NA\ 60\times$ oil immersion objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on your specimen?

Exercise:

Problem:

An amoeba is 0.305 cm away from the 0.300 cm focal length objective lens of a microscope. (a) Where is the image formed by the objective lens? (b) What is this image's magnification? (c) An eyepiece with a 2.00 cm focal length is placed 20.0 cm from the objective. Where is the final image? (d) What magnification is produced by the eyepiece? (e) What is the overall magnification? (See [link].)

Solution:

(a) +18.3 cm (on the eyepiece side of the objective lens)

(b) -60.0

(c) -11.3 cm (on the objective side of the eyepiece)

(d) +6.67

(e) -400

Exercise:

Problem:

You are using a standard microscope with a $0.10NA\ 4\times$ objective and switch to a $0.65NA\ 40\times$ objective. What are the acceptance angles for each? Compare and comment on the values. Which would you use first to locate the target area on of your specimen? (See [link].)

Exercise:

Problem: Unreasonable Results

Your friends show you an image through a microscope. They tell you that the microscope has an objective with a 0.500 cm focal length and an eyepiece with a 5.00 cm focal length. The resulting overall magnification is 250,000. Are these viable values for a microscope?

Glossary

compound microscope

a microscope constructed from two convex lenses, the first serving as the ocular lens(close to the eye) and the second serving as the objective lens

objective lens

the lens nearest to the object being examined

eyepiece

the lens or combination of lenses in an optical instrument nearest to the eye of the observer

numerical aperture

a number or measure that expresses the ability of a lens to resolve fine detail in an object being observed. Derived by mathematical formula

Equation:

$$\mathrm{NA} = n \sin \alpha,$$

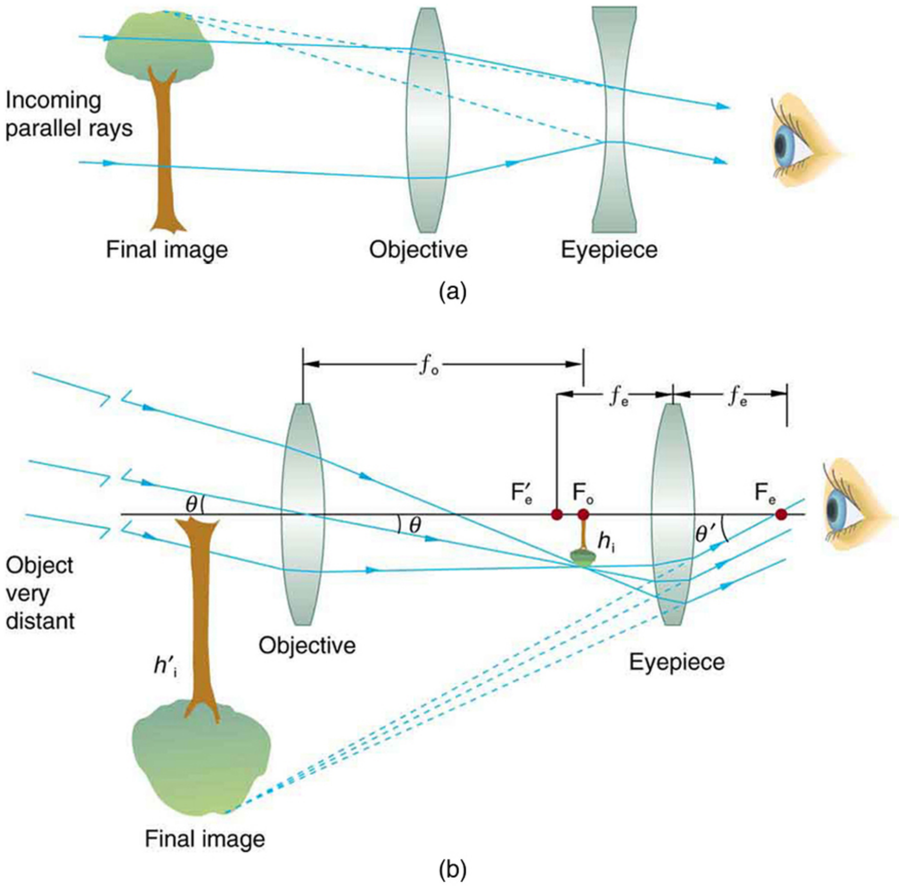
where n is the refractive index of the medium between the lens and the specimen and $\alpha = \theta/2$

Telescopes

- Outline the invention of a telescope.
- Describe the working of a telescope.

Telescopes are meant for viewing distant objects, producing an image that is larger than the image that can be seen with the unaided eye. Telescopes gather far more light than the eye, allowing dim objects to be observed with greater magnification and better resolution. Although Galileo is often credited with inventing the telescope, he actually did not. What he did was more important. He constructed several early telescopes, was the first to study the heavens with them, and made monumental discoveries using them. Among these are the moons of Jupiter, the craters and mountains on the Moon, the details of sunspots, and the fact that the Milky Way is composed of vast numbers of individual stars.

[link](a) shows a telescope made of two lenses, the convex objective and the concave eyepiece, the same construction used by Galileo. Such an arrangement produces an upright image and is used in spyglasses and opera glasses.



(a) Galileo made telescopes with a convex objective and a concave eyepiece. These produce an upright image and are used in spyglasses. (b) Most simple telescopes have two convex lenses. The objective forms a case 1 image that is the object for the eyepiece. The eyepiece forms a case 2 final image that is magnified.

The most common two-lens telescope, like the simple microscope, uses two convex lenses and is shown in [link](b). The object is so far away from the telescope that it is essentially at infinity compared with the focal lengths of the lenses ($d_{\text{o}} \approx \infty$). The first image is thus produced at $d_{\text{i}} = f_{\text{o}}$, as shown in the figure. To prove this, note that

Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{f_{\text{o}}} - \frac{1}{d_{\text{o}}} = \frac{1}{f_{\text{o}}} - \frac{1}{\infty}.$$

Because $1/\infty = 0$, this simplifies to
Equation:

$$\frac{1}{d_{\text{i}}} = \frac{1}{f_{\text{o}}},$$

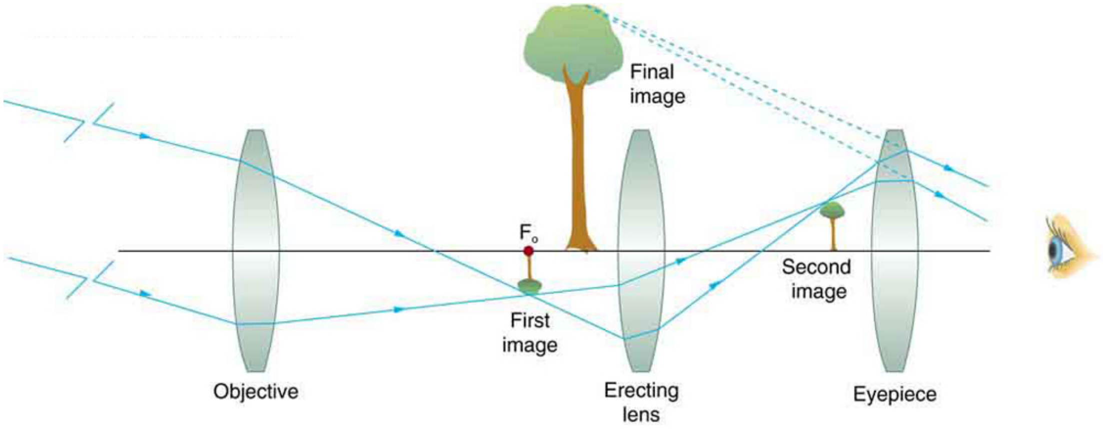
which implies that $d_{\text{i}} = f_{\text{o}}$, as claimed. It is true that for any distant object and any lens or mirror, the image is at the focal length.

The first image formed by a telescope objective as seen in [link](b) will not be large compared with what you might see by looking at the object directly. For example, the spot formed by sunlight focused on a piece of paper by a magnifying glass is the image of the Sun, and it is small. The telescope eyepiece (like the microscope eyepiece) magnifies this first image. The distance between the eyepiece and the objective lens is made slightly less than the sum of their focal lengths so that the first image is closer to the eyepiece than its focal length. That is, $d_{\text{o}}\prime$ is less than f_{e}, and so the eyepiece forms a case 2 image that is large and to the left for easy viewing. If the angle subtended by an object as viewed by the unaided eye is θ, and the angle subtended by the telescope image is $\theta\prime$, then the **angular magnification** M is defined to be their ratio. That is, $M = \theta\prime/\theta$. It can be shown that the angular magnification of a telescope is related to the focal lengths of the objective and eyepiece; and is given by
Equation:

$$M = \frac{\theta\prime}{\theta} = -\frac{f_{\text{o}}}{f_{\text{e}}}.$$

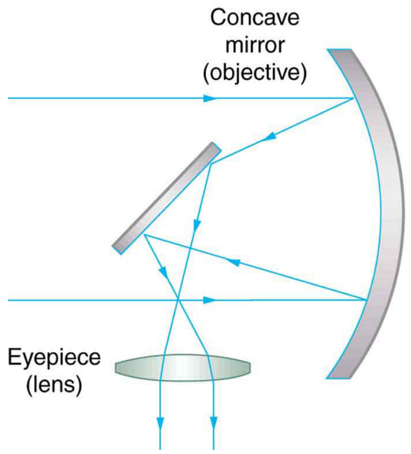
The minus sign indicates the image is inverted. To obtain the greatest angular magnification, it is best to have a long focal length objective and a short focal length eyepiece. The greater the angular magnification M, the larger an object will appear when viewed through a telescope, making more details visible. Limits to observable details are imposed by many factors, including lens quality and atmospheric disturbance.

The image in most telescopes is inverted, which is unimportant for observing the stars but a real problem for other applications, such as telescopes on ships or telescopic gun sights. If an upright image is needed, Galileo’s arrangement in [link](a) can be used. But a more common arrangement is to use a third convex lens as an eyepiece, increasing the distance between the first two and inverting the image once again as seen in [link].



This arrangement of three lenses in a telescope produces an upright final image. The first two lenses are far enough apart that the second lens inverts the image of the first one more time. The third lens acts as a magnifier and keeps the image upright and in a location that is easy to view.

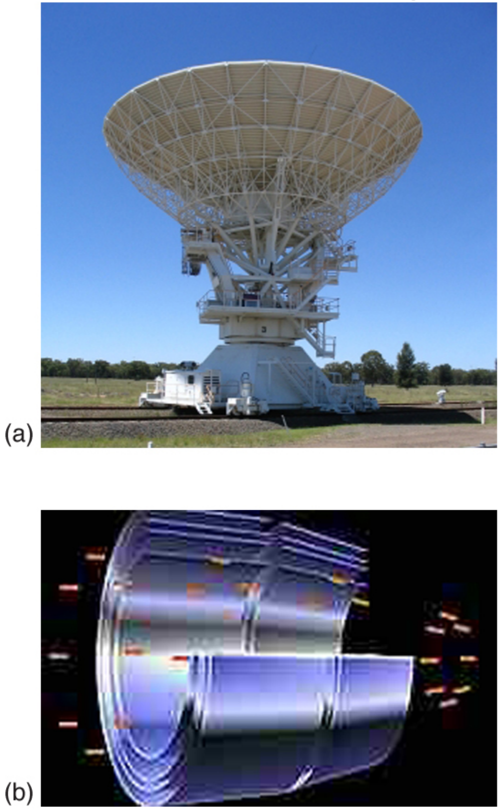
A telescope can also be made with a concave mirror as its first element or objective, since a concave mirror acts like a convex lens as seen in [link]. Flat mirrors are often employed in optical instruments to make them more compact or to send light to cameras and other sensing devices. There are many advantages to using mirrors rather than lenses for telescope objectives. Mirrors can be constructed much larger than lenses and can, thus, gather large amounts of light, as needed to view distant galaxies, for example. Large and relatively flat mirrors have very long focal lengths, so that great angular magnification is possible.



A two-element telescope composed of a mirror as the objective and a lens for the eyepiece is shown. This telescope forms an image in the same manner as the two-convex-lens telescope already discussed, but it does not suffer from chromatic aberrations. Such telescopes can gather more light, since larger mirrors than lenses can be constructed.

Telescopes, like microscopes, can utilize a range of frequencies from the electromagnetic spectrum. [link](a) shows the Australia Telescope Compact

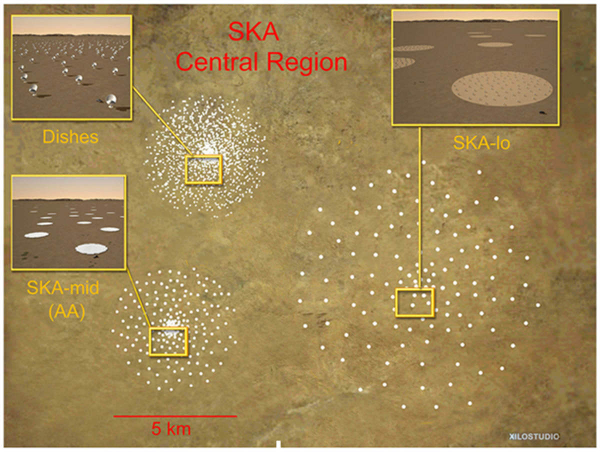
Array, which uses six 22-m antennas for mapping the southern skies using radio waves. [link](b) shows the focusing of x rays on the Chandra X-ray Observatory—a satellite orbiting earth since 1999 and looking at high temperature events as exploding stars, quasars, and black holes. X rays, with much more energy and shorter wavelengths than RF and light, are mainly absorbed and not reflected when incident perpendicular to the medium. But they can be reflected when incident at small glancing angles, much like a rock will skip on a lake if thrown at a small angle. The mirrors for the Chandra consist of a long barrelled pathway and 4 pairs of mirrors to focus the rays at a point 10 meters away from the entrance. The mirrors are extremely smooth and consist of a glass ceramic base with a thin coating of metal (iridium). Four pairs of precision manufactured mirrors are exquisitely shaped and aligned so that x rays ricochet off the mirrors like bullets off a wall, focusing on a spot.



(a) The Australia Telescope Compact Array at Narrabri (500 km NW of Sydney). (credit: Ian

Bailey) (b) The focusing of x rays on the Chandra Observatory, a satellite orbiting earth. X rays ricochet off 4 pairs of mirrors forming a barrelled pathway leading to the focus point. (credit: NASA)

A current exciting development is a collaborative effort involving 17 countries to construct a Square Kilometre Array (SKA) of telescopes capable of covering from 80 MHz to 2 GHz. The initial stage of the project is the construction of the Australian Square Kilometre Array Pathfinder in Western Australia (see [link]). The project will use cutting-edge technologies such as **adaptive optics** in which the lens or mirror is constructed from lots of carefully aligned tiny lenses and mirrors that can be manipulated using computers. A range of rapidly changing distortions can be minimized by deforming or tilting the tiny lenses and mirrors. The use of adaptive optics in vision correction is a current area of research.



An artist’s impression of the Australian Square Kilometre Array Pathfinder in Western

Australia is displayed. (credit: SPDO, XILOSTUDIOS)

Section Summary

- Simple telescopes can be made with two lenses. They are used for viewing objects at large distances and utilize the entire range of the electromagnetic spectrum.
- The angular magnification M for a telescope is given by **Equation:**

  $$M=\frac{\theta\prime}{\theta}=-\frac{f_{\mathrm{o}}}{f_{\mathrm{e}}},$$

  where θ is the angle subtended by an object viewed by the unaided eye, $\theta\prime$ is the angle subtended by a magnified image, and f_{o} and f_{e} are the focal lengths of the objective and the eyepiece.

Conceptual Questions

Exercise:

Problem:

If you want your microscope or telescope to project a real image onto a screen, how would you change the placement of the eyepiece relative to the objective?

Problem Exercises

Unless otherwise stated, the lens-to-retina distance is 2.00 cm.
Exercise:

Problem:

What is the angular magnification of a telescope that has a 100 cm focal length objective and a 2.50 cm focal length eyepiece?

Solution:

-40.0

Exercise:

Problem:

Find the distance between the objective and eyepiece lenses in the telescope in the above problem needed to produce a final image very far from the observer, where vision is most relaxed. Note that a telescope is normally used to view very distant objects.

Exercise:

Problem:

A large reflecting telescope has an objective mirror with a 10.0 m radius of curvature. What angular magnification does it produce when a 3.00 m focal length eyepiece is used?

Solution:

-1.67

Exercise:

Problem:

A small telescope has a concave mirror with a 2.00 m radius of curvature for its objective. Its eyepiece is a 4.00 cm focal length lens. (a) What is the telescope’s angular magnification? (b) What angle is subtended by a 25,000 km diameter sunspot? (c) What is the angle of its telescopic image?

Exercise:

Problem:

A $7.5\times$ binocular produces an angular magnification of -7.50, acting like a telescope. (Mirrors are used to make the image upright.) If the binoculars have objective lenses with a 75.0 cm focal length, what is the focal length of the eyepiece lenses?

Solution:

$+10.0\text{ cm}$

Exercise:

Problem: Construct Your Own Problem

Consider a telescope of the type used by Galileo, having a convex objective and a concave eyepiece as illustrated in [link](a). Construct a problem in which you calculate the location and size of the image produced. Among the things to be considered are the focal lengths of the lenses and their relative placements as well as the size and location of the object. Verify that the angular magnification is greater than one. That is, the angle subtended at the eye by the image is greater than the angle subtended by the object.

Glossary

adaptive optics
: optical technology in which computers adjust the lenses and mirrors in a device to correct for image distortions

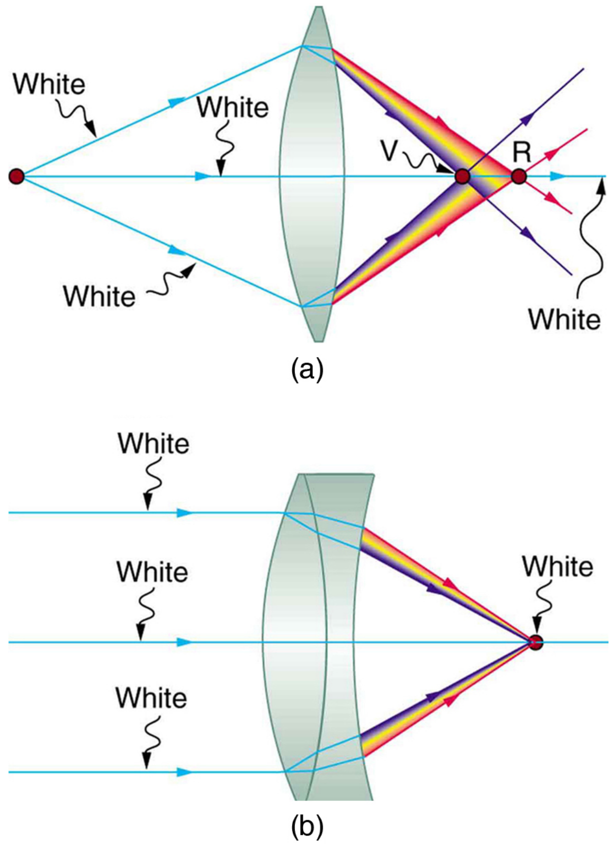
angular magnification
: a ratio related to the focal lengths of the objective and eyepiece and given as $M=-\frac{f_{\text{o}}}{f_{\text{e}}}$

Aberrations

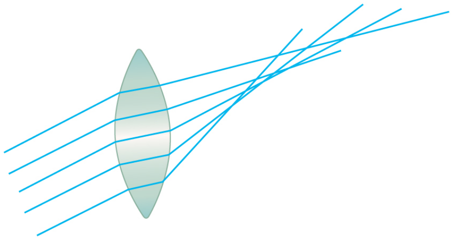
- Describe optical aberration.

Real lenses behave somewhat differently from how they are modeled using the thin lens equations, producing **aberrations**. An aberration is a distortion in an image. There are a variety of aberrations due to a lens size, material, thickness, and position of the object. One common type of aberration is chromatic aberration, which is related to color. Since the index of refraction of lenses depends on color or wavelength, images are produced at different places and with different magnifications for different colors. (The law of reflection is independent of wavelength, and so mirrors do not have this problem. This is another advantage for mirrors in optical systems such as telescopes.) [link](a) shows chromatic aberration for a single convex lens and its partial correction with a two-lens system. Violet rays are bent more than red, since they have a higher index of refraction and are thus focused closer to the lens. The diverging lens partially corrects this, although it is usually not possible to do so completely. Lenses of different materials and having different dispersions may be used. For example an achromatic doublet consisting of a converging lens made of crown glass and a diverging lens made of flint glass in contact can dramatically reduce chromatic aberration (see [link](b)).

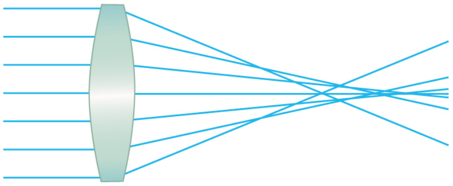
Quite often in an imaging system the object is off-center. Consequently, different parts of a lens or mirror do not refract or reflect the image to the same point. This type of aberration is called a coma and is shown in [link]. The image in this case often appears pear-shaped. Another common aberration is spherical aberration where rays converging from the outer edges of a lens converge to a focus closer to the lens and rays closer to the axis focus further (see [link]). Aberrations due to astigmatism in the lenses of the eyes are discussed in Vision Correction, and a chart used to detect astigmatism is shown in [link]. Such aberrations and can also be an issue with manufactured lenses.



(a) Chromatic aberration is caused by the dependence of a lens’s index of refraction on color (wavelength). The lens is more powerful for violet (V) than for red (R), producing images with different locations and magnifications. (b) Multiple-lens systems can partially correct chromatic aberrations, but they may require lenses of different materials and add to the expense of optical systems such as cameras.



A coma is an aberration caused by an object that is off-center, often resulting in a pear-shaped image. The rays originate from points that are not on the optical axis and they do not converge at one common focal point.



Spherical aberration is caused by rays focusing at different distances from the lens.

The image produced by an optical system needs to be bright enough to be discerned. It is often a challenge to obtain a sufficiently bright image. The brightness is determined by the amount of light passing through the optical system. The optical components determining the brightness are the diameter of the lens and the diameter of pupils, diaphragms or aperture stops placed in front of lenses. Optical systems often have entrance and exit pupils to specifically reduce aberrations but they inevitably reduce brightness as well. Consequently, optical systems need to strike a balance between the various components used. The iris in the eye dilates and constricts, acting as an entrance pupil. You can see objects more clearly by looking through a small hole made with your hand in the shape of a fist. Squinting, or using a small hole in a piece of paper, also will make the object sharper.

So how are aberrations corrected? The lenses may also have specially shaped surfaces, as opposed to the simple spherical shape that is relatively easy to produce. Expensive camera lenses are large in diameter, so that they can gather more light, and need several elements to correct for various aberrations. Further, advances in materials science have resulted in lenses with a range of refractive indices—technically referred to as graded index (GRIN) lenses. Spectacles often have the ability to provide a range of focusing ability using similar techniques. GRIN lenses are particularly important at the end of optical fibers in endoscopes. Advanced computing techniques allow for a range of corrections on images after the image has been collected and certain characteristics of the optical system are known. Some of these techniques are sophisticated versions of what are available on commercial packages like Adobe Photoshop.

Section Summary

- Aberrations or image distortions can arise due to the finite thickness of optical instruments, imperfections in the optical components, and limitations on the ways in which the components are used.
- The means for correcting aberrations range from better components to computational techniques.

Conceptual Questions

Exercise:

Problem:

List the various types of aberrations. What causes them and how can each be reduced?

Problem Exercises

Exercise:

Problem: **Integrated Concepts**

(a) During laser vision correction, a brief burst of 193 nm ultraviolet light is projected onto the cornea of the patient. It makes a spot 1.00 mm in diameter and deposits 0.500 mJ of energy. Calculate the depth of the layer ablated, assuming the corneal tissue has the same properties as water and is initially at º . The tissue’s temperature is increased to º and evaporated without further temperature increase.

(b) Does your answer imply that the shape of the cornea can be finely controlled?

Solution:

(a) μ

(b) Yes, this thickness implies that the shape of the cornea can be very finely controlled, producing normal distant vision in more than 90% of patients.

Glossary

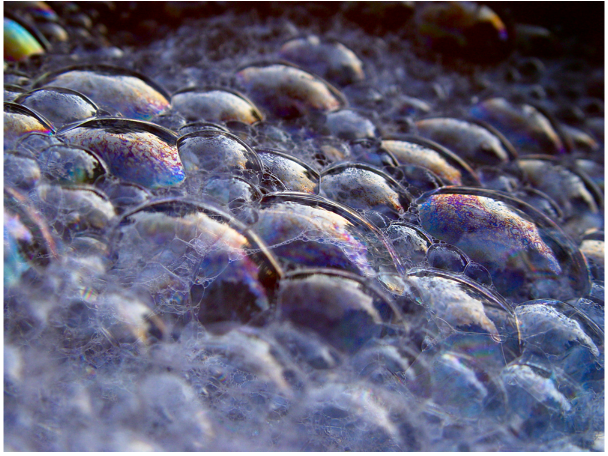
aberration
failure of rays to converge at one focus because of limitations or defects in a lens or mirror

Introduction to Wave Optics class="introduction"

The colors reflected by this compact disc vary with angle and are not caused by pigments. Colors such as these are direct evidence of the wave character of light. (credit: Infopro, Wikimedia Commons)

Examine a compact disc under white light, noting the colors observed and locations of the colors. Determine if the spectra are formed by diffraction from circular lines centered at the middle of the disc and, if so, what is their spacing. If not, determine the type of spacing. Also with the CD, explore the spectra of a few light sources, such as a candle flame, incandescent bulb, halogen light, and fluorescent light. Knowing the spacing of the rows of pits in the compact disc, estimate the maximum spacing that will allow the given number of megabytes of information to be stored.

If you have ever looked at the reds, blues, and greens in a sunlit soap bubble and wondered how straw-colored soapy water could produce them, you have hit upon one of the many phenomena that can only be explained by the wave character of light (see [link]). The same is true for the colors seen in an oil slick or in the light reflected from a compact disc. These and other interesting phenomena, such as the dispersion of white light into a rainbow of colors when passed through a narrow slit, cannot be explained fully by geometric optics. In these cases, light interacts with small objects and exhibits its wave characteristics. The branch of optics that considers the behavior of light when it exhibits wave characteristics (particularly when it interacts with small objects) is called wave optics (sometimes called physical optics). It is the topic of this chapter.



These soap bubbles exhibit brilliant colors when exposed to sunlight. How are the colors produced if they are not pigments in the soap? (credit: Scott Robinson, Flickr)

The Wave Aspect of Light: Interference

- Discuss the wave character of light.
- Identify the changes when light enters a medium.

We know that visible light is the type of electromagnetic wave to which our eyes respond. Like all other electromagnetic waves, it obeys the equation

Equation:

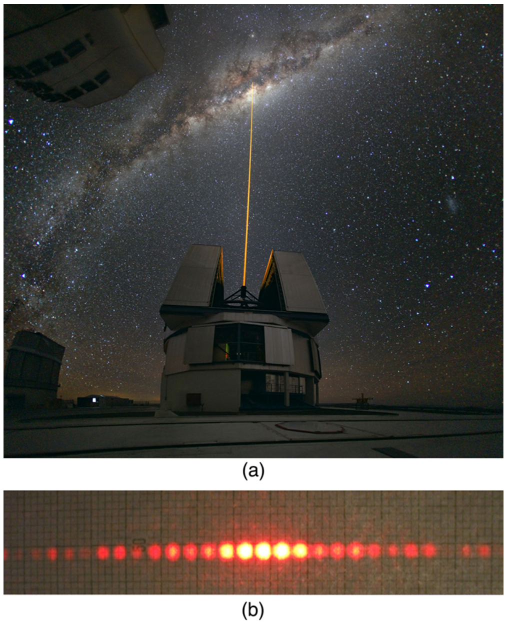
$$c = f\lambda,$$

where $c = 3 \times 10^8\ \text{m/s}$ is the speed of light in vacuum, f is the frequency of the electromagnetic waves, and λ is its wavelength. The range of visible wavelengths is approximately 380 to 760 nm. As is true for all waves, light travels in straight lines and acts like a ray when it interacts with objects several times as large as its wavelength. However, when it interacts with smaller objects, it displays its wave characteristics prominently. Interference is the hallmark of a wave, and in [link] both the ray and wave characteristics of light can be seen. The laser beam emitted by the observatory epitomizes a ray, traveling in a straight line. However, passing a pure-wavelength beam through vertical slits with a size close to the wavelength of the beam reveals the wave character of light, as the beam spreads out horizontally into a pattern of bright and dark regions caused by systematic constructive and destructive interference. Rather than spreading out, a ray would continue traveling straight ahead after passing through slits.

Note:

Making Connections: Waves

The most certain indication of a wave is interference. This wave characteristic is most prominent when the wave interacts with an object that is not large compared with the wavelength. Interference is observed for water waves, sound waves, light waves, and (as we will see in Special Relativity) for matter waves, such as electrons scattered from a crystal.



(a) The laser beam emitted by an observatory acts like a ray, traveling in a straight line. This laser beam is from the Paranal Observatory of the European Southern Observatory. (credit: Yuri Beletsky, European Southern Observatory) (b) A laser beam passing through a grid of vertical slits produces an interference pattern—characteristic of a wave. (credit: Shim'on and Slava Rybka, Wikimedia Commons)

Light has wave characteristics in various media as well as in a vacuum. When light goes from a vacuum to some medium, like water, its speed and wavelength change, but its frequency f remains the same. (We can think of light as a forced oscillation that must have the frequency of the original source.) The speed of light in a medium is $v = c/n$, where n is its index of refraction. If we divide both sides of equation $c = f\lambda$ by n, we get $c/n = v = f\lambda/n$. This implies that $v = f\lambda_{\text{n}}$, where λ_{n} is the **wavelength in a medium** and that

Equation:

$$\lambda_{\text{n}} = \frac{\lambda}{n},$$

where λ is the wavelength in vacuum and n is the medium's index of refraction. Therefore, the wavelength of light is smaller in any medium than it is in vacuum. In water, for example, which has $n = 1.333$, the range of visible wavelengths is $(380\text{ nm})/1.333$ to $(760\text{ nm})/1.333$, or $\lambda_{\text{n}} = 285\text{ to }570\text{ nm}$. Although wavelengths change while traveling from one medium to another, colors do not, since colors are associated with frequency.

Section Summary

- Wave optics is the branch of optics that must be used when light interacts with small objects or whenever the wave characteristics of light are considered.
- Wave characteristics are those associated with interference and diffraction.
- Visible light is the type of electromagnetic wave to which our eyes respond and has a wavelength in the range of 380 to 760 nm.
- Like all EM waves, the following relationship is valid in vacuum: $c = f\lambda$, where $c = 3 \times 10^{8}\text{ m/s}$ is the speed of light, f is the frequency of the electromagnetic wave, and λ is its wavelength in vacuum.
- The wavelength λ_{n} of light in a medium with index of refraction n is $\lambda_{\text{n}} = \lambda/n$. Its frequency is the same as in vacuum.

Conceptual Questions

Exercise:

Problem:

What type of experimental evidence indicates that light is a wave?

Exercise:

Problem:

Give an example of a wave characteristic of light that is easily observed outside the laboratory.

Problems & Exercises

Exercise:

Problem:

Show that when light passes from air to water, its wavelength decreases to 0.750 times its original value.

Solution:

$1/1.333 = 0.750$

Exercise:

Problem:

Find the range of visible wavelengths of light in crown glass.

Exercise:

Problem:

What is the index of refraction of a material for which the wavelength of light is 0.671 times its value in a vacuum? Identify the likely substance.

Solution:

1.49, Polystyrene

Exercise:

Problem:

Analysis of an interference effect in a clear solid shows that the wavelength of light in the solid is 329 nm. Knowing this light comes from a He-Ne laser and has a wavelength of 633 nm in air, is the substance zircon or diamond?

Exercise:

Problem:

What is the ratio of thicknesses of crown glass and water that would contain the same number of wavelengths of light?

Solution:

0.877 glass to water

Glossary

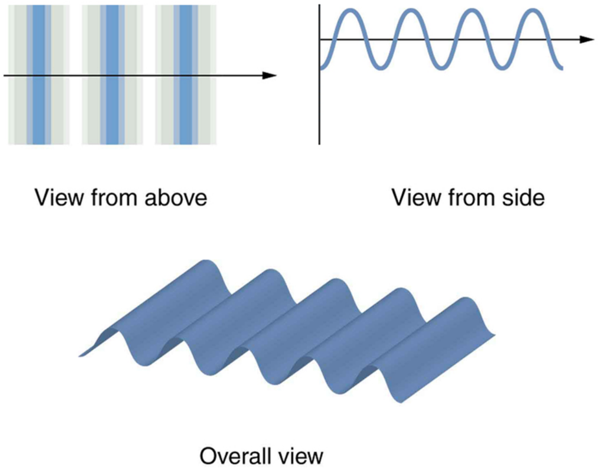
wavelength in a medium

$\lambda_{\text{n}} = \lambda/n$, where λ is the wavelength in vacuum, and n is the index of refraction of the medium

Huygens's Principle: Diffraction

- Discuss the propagation of transverse waves.
- Discuss Huygens’s principle.
- Explain the bending of light.

[link] shows how a transverse wave looks as viewed from above and from the side. A light wave can be imagined to propagate like this, although we do not actually see it wiggling through space. From above, we view the wavefronts (or wave crests) as we would by looking down on the ocean waves. The side view would be a graph of the electric or magnetic field. The view from above is perhaps the most useful in developing concepts about wave optics.



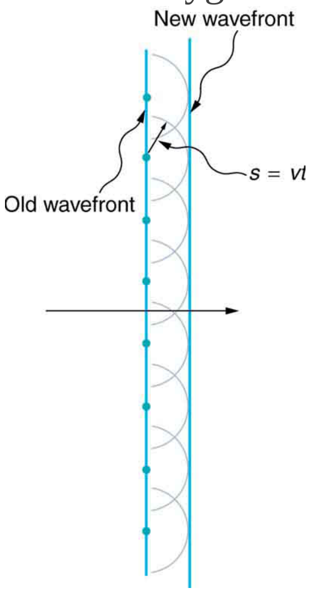
A transverse wave, such as an electromagnetic wave like light, as viewed from above and from the side. The direction of propagation is perpendicular to the wavefronts (or wave crests) and is represented by an arrow like a ray.

The Dutch scientist Christiaan Huygens (1629–1695) developed a useful technique for determining in detail how and where waves propagate.

Starting from some known position, **Huygens's principle** states that:

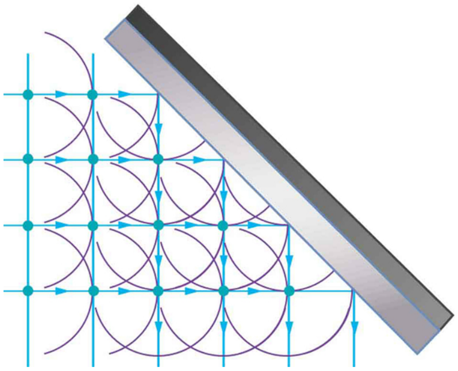
Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.

[link] shows how Huygens's principle is applied. A wavefront is the long edge that moves, for example, the crest or the trough. Each point on the wavefront emits a semicircular wave that moves at the propagation speed v. These are drawn at a time t later, so that they have moved a distance $s = \text{vt}$. The new wavefront is a line tangent to the wavelets and is where we would expect the wave to be a time t later. Huygens's principle works for all types of waves, including water waves, sound waves, and light waves. We will find it useful not only in describing how light waves propagate, but also in explaining the laws of reflection and refraction. In addition, we will see that Huygens's principle tells us how and where light rays interfere.



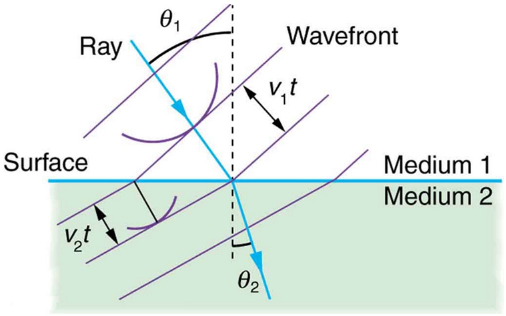
Huygens's principle applied to a straight wavefront. Each point on the wavefront emits a semicircular wavelet that moves a distance $s=vt$. The new wavefront is a line tangent to the wavelets.

[link] shows how a mirror reflects an incoming wave at an angle equal to the incident angle, verifying the law of reflection. As the wavefront strikes the mirror, wavelets are first emitted from the left part of the mirror and then the right. The wavelets closer to the left have had time to travel farther, producing a wavefront traveling in the direction shown.



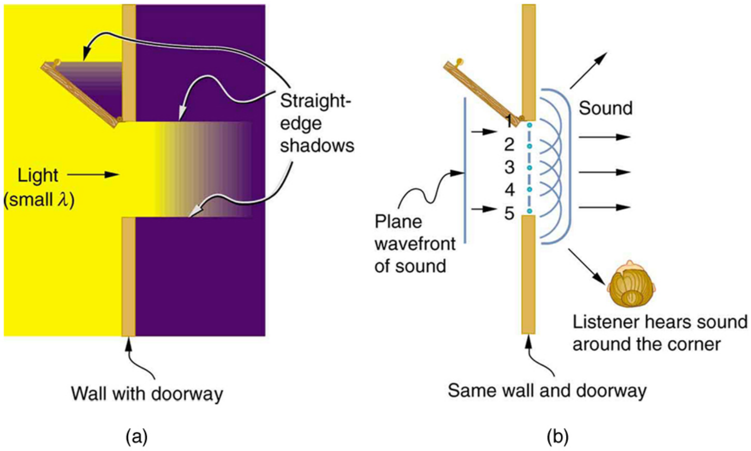
Huygens’s principle applied to a straight wavefront striking a mirror. The wavelets shown were emitted as each point on the wavefront struck the mirror. The tangent to these wavelets shows that the new wavefront has been reflected at an angle equal to the incident angle. The direction of propagation is perpendicular to the wavefront, as shown by the downward-pointing arrows.

The law of refraction can be explained by applying Huygens’s principle to a wavefront passing from one medium to another (see [link]). Each wavelet in the figure was emitted when the wavefront crossed the interface between the media. Since the speed of light is smaller in the second medium, the waves do not travel as far in a given time, and the new wavefront changes direction as shown. This explains why a ray changes direction to become closer to the perpendicular when light slows down. Snell’s law can be derived from the geometry in [link], but this is left as an exercise for ambitious readers.



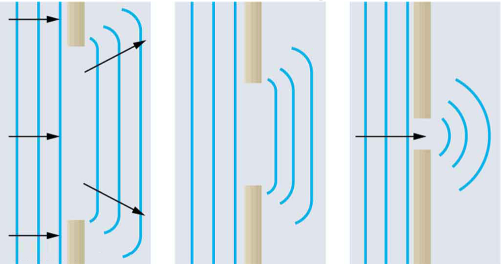
Huygens’s principle applied to a straight wavefront traveling from one medium to another where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

What happens when a wave passes through an opening, such as light shining through an open door into a dark room? For light, we expect to see a sharp shadow of the doorway on the floor of the room, and we expect no light to bend around corners into other parts of the room. When sound passes through a door, we expect to hear it everywhere in the room and, thus, expect that sound spreads out when passing through such an opening (see [link]). What is the difference between the behavior of sound waves and light waves in this case? The answer is that light has very short wavelengths and acts like a ray. Sound has wavelengths on the order of the size of the door and bends around corners (for frequency of 1000 Hz, $\lambda = c/f = (330 \text{ m/s})/(1000 \text{ s}^{-1}) = 0.33 \text{ m}$, about three times smaller than the width of the doorway).



(a) Light passing through a doorway makes a sharp outline on the floor. Since light’s wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the door.

If we pass light through smaller openings, often called slits, we can use Huygens’s principle to see that light bends as sound does (see [link]). The bending of a wave around the edges of an opening or an obstacle is called **diffraction**. Diffraction is a wave characteristic and occurs for all types of waves. If diffraction is observed for some phenomenon, it is evidence that the phenomenon is a wave. Thus the horizontal diffraction of the laser beam after it passes through slits in [link] is evidence that light is a wave.



Huygens’s principle applied to a straight wavefront striking an opening. The edges of the wavefront bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

Section Summary

- An accurate technique for determining how and where waves propagate is given by Huygens's principle: Every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets.
- Diffraction is the bending of a wave around the edges of an opening or other obstacle.

Conceptual Questions

Exercise:

Problem:

How do wave effects depend on the size of the object with which the wave interacts? For example, why does sound bend around the corner of a building while light does not?

Exercise:

Problem:

Under what conditions can light be modeled like a ray? Like a wave?

Exercise:

Problem:

Go outside in the sunlight and observe your shadow. It has fuzzy edges even if you do not. Is this a diffraction effect? Explain.

Exercise:

Problem:

Why does the wavelength of light decrease when it passes from vacuum into a medium? State which attributes change and which stay the same and, thus, require the wavelength to decrease.

Exercise:

Problem: Does Huygens’s principle apply to all types of waves?

Glossary

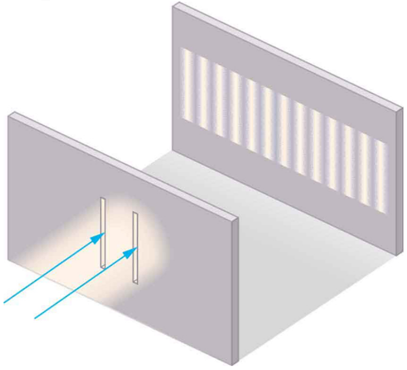
diffraction
the bending of a wave around the edges of an opening or an obstacle

Huygens’s principle
every point on a wavefront is a source of wavelets that spread out in the forward direction at the same speed as the wave itself. The new wavefront is a line tangent to all of the wavelets

Young’s Double Slit Experiment

- Explain the phenomena of interference.
- Define constructive interference for a double slit and destructive interference for a double slit.

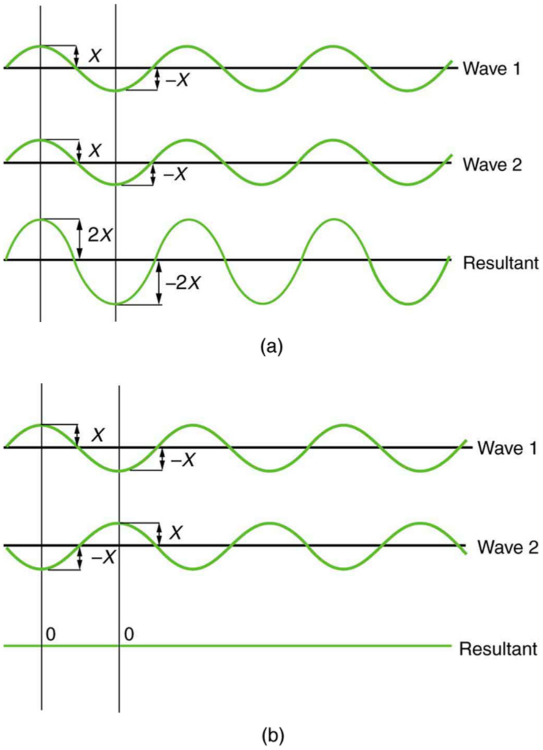
Although Christiaan Huygens thought that light was a wave, Isaac Newton did not. Newton felt that there were other explanations for color, and for the interference and diffraction effects that were observable at the time. Owing to Newton’s tremendous stature, his view generally prevailed. The fact that Huygens’s principle worked was not considered evidence that was direct enough to prove that light is a wave. The acceptance of the wave character of light came many years later when, in 1801, the English physicist and physician Thomas Young (1773–1829) did his now-classic double slit experiment (see [link]).



Young’s double slit experiment. Here pure-wavelength light sent through a pair of vertical slits is diffracted into a pattern on the screen of numerous vertical lines spread out horizontally. Without diffraction and interference, the light would

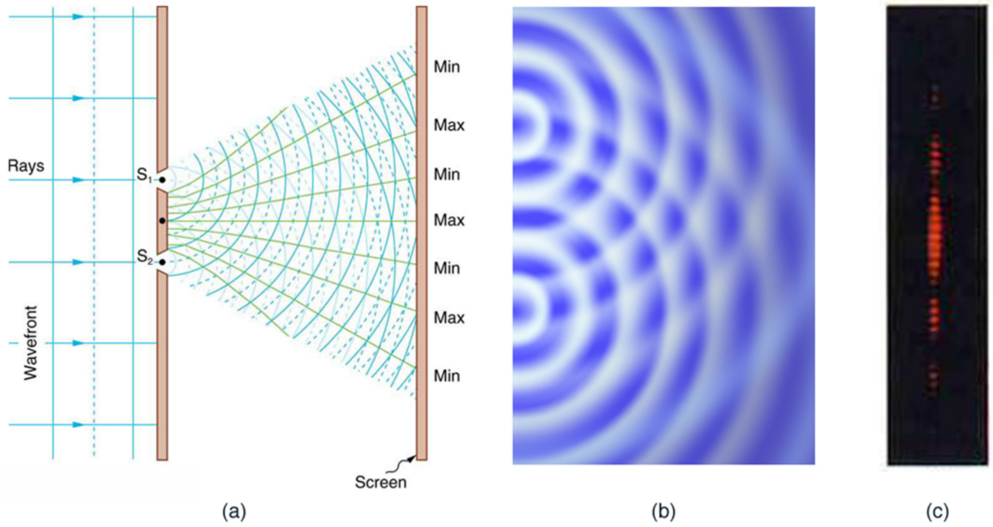
simply make two
lines on the screen.

Why do we not ordinarily observe wave behavior for light, such as observed in Young’s double slit experiment? First, light must interact with something small, such as the closely spaced slits used by Young, to show pronounced wave effects. Furthermore, Young first passed light from a single source (the Sun) through a single slit to make the light somewhat coherent. By **coherent**, we mean waves are in phase or have a definite phase relationship. **Incoherent** means the waves have random phase relationships. Why did Young then pass the light through a double slit? The answer to this question is that two slits provide two coherent light sources that then interfere constructively or destructively. Young used sunlight, where each wavelength forms its own pattern, making the effect more difficult to see. We illustrate the double slit experiment with monochromatic (single λ) light to clarify the effect. [link] shows the pure constructive and destructive interference of two waves having the same wavelength and amplitude.



The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

When light passes through narrow slits, it is diffracted into semicircular waves, as shown in [link](a). Pure constructive interference occurs where the waves are crest to crest or trough to trough. Pure destructive interference occurs where they are crest to trough. The light must fall on a screen and be scattered into our eyes for us to see the pattern. An analogous pattern for water waves is shown in [link](b). Note that regions of constructive and destructive interference move out from the slits at well-defined angles to the original beam. These angles depend on wavelength and the distance between the slits, as we shall see below.



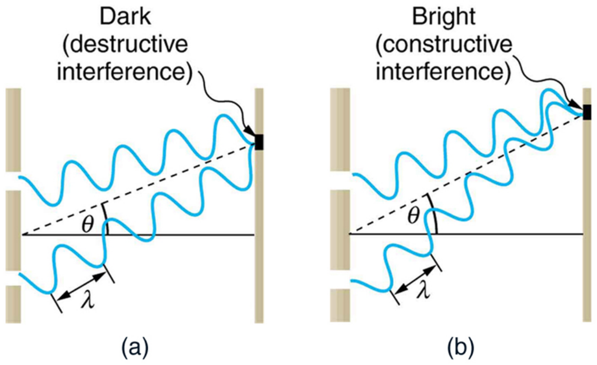
Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto a screen and is scattered into our eyes. (b) Double slit interference pattern for water waves are nearly identical to that for light. Wave action is greatest in regions of constructive interference and least in regions of destructive interference. (c) When light that has passed through double slits falls on a screen, we see a pattern such as this. (credit: PASCO)

To understand the double slit interference pattern, we consider how two waves travel from the slits to the screen, as illustrated in [link]. Each slit is a different distance from a given point on the screen. Thus different numbers of wavelengths fit into each path. Waves start out from the slits in phase (crest to crest), but they may end up out of phase (crest to trough) at the screen if the paths differ in length by half a wavelength, interfering destructively as shown in [link](a). If the paths differ by a whole wavelength, then the waves arrive in phase (crest to crest) at the screen, interfering constructively as shown in [link](b). More generally, if the paths taken by the two waves differ by any half-integral number of wavelengths [$(1/2)\lambda$, $(3/2)\lambda$, $(5/2)\lambda$, etc.], then destructive interference occurs. Similarly, if the paths taken by the two waves differ by any integral number of wavelengths (λ, 2λ, 3λ, etc.), then constructive interference occurs.

Note:

Take-Home Experiment: Using Fingers as Slits

Look at a light, such as a street lamp or incandescent bulb, through the narrow gap between two fingers held close together. What type of pattern do you see? How does it change when you allow the fingers to move a little farther apart? Is it more distinct for a monochromatic source, such as the yellow light from a sodium vapor lamp, than for an incandescent bulb?



Waves follow different paths from the slits to a common point on a screen. (a) Destructive interference occurs here, because one path is a half wavelength longer than the other. The waves start in phase but arrive out of phase. (b) Constructive interference occurs here because one path is a whole wavelength longer than the other. The waves start out and arrive in phase.

[link] shows how to determine the path length difference for waves traveling from two slits to a common point on a screen. If the screen is a large distance away compared with the distance between the slits, then the angle θ between the path and a line from the slits to the screen (see the figure) is nearly the same for each path. The difference between the paths is shown in the figure; simple trigonometry shows it to be d sin θ, where d is the distance between the slits. To obtain **constructive interference for a double slit**, the path length difference must be an integral multiple of the wavelength, or

Equation:

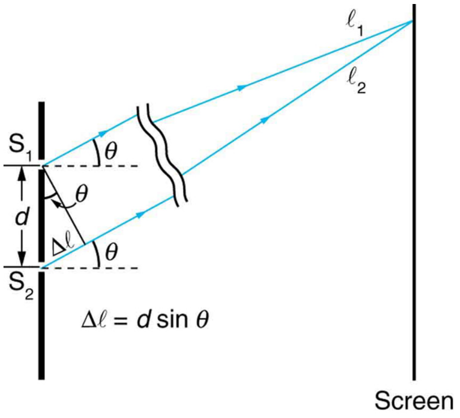
$$d\ \sin\theta = m\lambda,\ \text{for}\ m = 0,\ 1,\ -1,\ 2,\ -2,\ \ldots\ \ (\text{constructive}).$$

Similarly, to obtain **destructive interference for a double slit**, the path length difference must be a half-integral multiple of the wavelength, or

Equation:

$$d\ \sin\theta = \left(m + \frac{1}{2}\right)\lambda,\ \text{for}\ m = 0,\ 1,\ -1,\ 2,\ -2,\ \ldots\ \ \text{(destructive)},$$

where λ is the wavelength of the light, d is the distance between slits, and θ is the angle from the original direction of the beam as discussed above. We call m the **order** of the interference. For example, $m = 4$ is fourth-order interference.



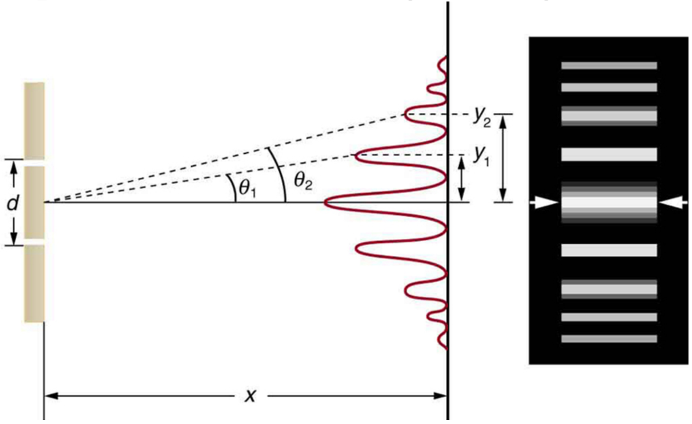
The paths from each slit to a common point on the screen differ by an amount d sin θ, assuming the distance to the screen is much greater than the distance between slits (not to scale here).

The equations for double slit interference imply that a series of bright and dark lines are formed. For vertical slits, the light spreads out horizontally on either side of the incident beam into a pattern called interference fringes, illustrated in [link]. The intensity of the bright fringes falls off on either side, being brightest at the center. The closer the slits are, the more is the spreading of the bright fringes. We can see this by examining the equation

Equation:

$$d\ \sin\theta = m\lambda, \text{ for } m = 0, 1, \ -1, 2, \ -2, \ \dots .$$

For fixed λ and m, the smaller d is, the larger θ must be, since $\sin\theta = m\lambda/\ d$. This is consistent with our contention that wave effects are most noticeable when the object the wave encounters (here, slits a distance d apart) is small. Small d gives large θ, hence a large effect.



The interference pattern for a double slit has an intensity that falls off with angle. The photograph shows multiple bright and dark lines, or fringes, formed by light passing through a double slit.

Example:
Finding a Wavelength from an Interference Pattern

Suppose you pass light from a He-Ne laser through two slits separated by 0.0100 mm and find that the third bright line on a screen is formed at an angle of 10.95º relative to the incident beam. What is the wavelength of the light?

Strategy

The third bright line is due to third-order constructive interference, which means that $m = 3$. We are given $d = 0.0100$ mm and $\theta = 10.95º$. The wavelength can thus be found using the equation $d \ \sin\theta = m\lambda$ for constructive interference.

Solution

The equation is $d \ \sin\theta = m\lambda$. Solving for the wavelength λ gives

Equation:

$$\lambda = \frac{d \ \sin\theta}{m}.$$

Substituting known values yields

Equation:

$$\begin{aligned}\lambda &= \frac{(0.0100 \text{ mm})(\sin 10.95º)}{3} \\ &= 6.33 \times 10^{-4} \text{ mm} = 633 \text{ nm}.\end{aligned}$$

Discussion

To three digits, this is the wavelength of light emitted by the common He-Ne laser. Not by coincidence, this red color is similar to that emitted by neon lights. More important, however, is the fact that interference patterns can be used to measure wavelength. Young did this for visible wavelengths. This analytical technique is still widely used to measure electromagnetic spectra. For a given order, the angle for constructive interference increases with λ, so that spectra (measurements of intensity versus wavelength) can be obtained.

Example:

Calculating Highest Order Possible

Interference patterns do not have an infinite number of lines, since there is a limit to how big m can be. What is the highest-order constructive interference possible with the system described in the preceding example?

Strategy and Concept

The equation $d\ \sin\theta = m\lambda$ (for $m = 0, 1, -1, 2, -2, \ldots$) describes constructive interference. For fixed values of d and λ, the larger m is, the larger sin θ is. However, the maximum value that sin θ can have is 1, for an angle of 90º. (Larger angles imply that light goes backward and does not reach the screen at all.) Let us find which m corresponds to this maximum diffraction angle.

Solution

Solving the equation $d\sin\theta = m\lambda$ for m gives

Equation:

$$m = \frac{d\sin\theta}{\lambda}.$$

Taking $\sin\theta = 1$ and substituting the values of d and λ from the preceding example gives

Equation:

$$m = \frac{(0.0100\text{ mm})(1)}{633\text{ nm}} \approx 15.8.$$

Therefore, the largest integer m can be is 15, or

Equation:

$$m = 15.$$

Discussion

The number of fringes depends on the wavelength and slit separation. The number of fringes will be very large for large slit separations. However, if the slit separation becomes much greater than the wavelength, the intensity of the interference pattern changes so that the screen has two bright lines cast by the slits, as expected when light behaves like a ray. We also note that the fringes get fainter further away from the center. Consequently, not all 15 fringes may be observable.

Section Summary

- Young’s double slit experiment gave definitive proof of the wave character of light.
- An interference pattern is obtained by the superposition of light from two slits.
- There is constructive interference when $d \sin \theta = m\lambda \; (\text{for } m = 0, 1, -1, 2, -2, \ldots)$, where d is the distance between the slits, θ is the angle relative to the incident direction, and m is the order of the interference.
- There is destructive interference when $d \;\sin \theta = \; m + \frac{1}{2} \; \lambda \; (\text{for } \; m = 0, 1, -1, 2, -2, \ldots)$.

Conceptual Questions

Exercise:

Problem:

Young’s double slit experiment breaks a single light beam into two sources. Would the same pattern be obtained for two independent sources of light, such as the headlights of a distant car? Explain.

Exercise:

Problem:

Suppose you use the same double slit to perform Young’s double slit experiment in air and then repeat the experiment in water. Do the angles to the same parts of the interference pattern get larger or smaller? Does the color of the light change? Explain.

Exercise:

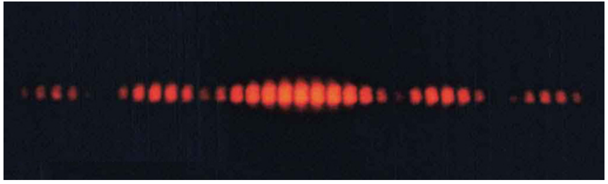
Problem:

Is it possible to create a situation in which there is only destructive interference? Explain.

Exercise:

Problem:

[link] shows the central part of the interference pattern for a pure wavelength of red light projected onto a double slit. The pattern is actually a combination of single slit and double slit interference. Note that the bright spots are evenly spaced. Is this a double slit or single slit characteristic? Note that some of the bright spots are dim on either side of the center. Is this a single slit or double slit characteristic? Which is smaller, the slit width or the separation between slits? Explain your responses.



This double slit interference pattern also shows signs of single slit interference. (credit: PASCO)

Problems & Exercises

Exercise:

Problem:

At what angle is the first-order maximum for 450-nm wavelength blue light falling on double slits separated by 0.0500 mm?

Solution:

0.516º

Exercise:

Problem:

Calculate the angle for the third-order maximum of 580-nm wavelength yellow light falling on double slits separated by 0.100 mm.

Exercise:

Problem:

What is the separation between two slits for which 610-nm orange light has its first maximum at an angle of 30.0^{o}?

Solution:

$1.22 \times 10^{-6}\ \text{m}$

Exercise:

Problem:

Find the distance between two slits that produces the first minimum for 410-nm violet light at an angle of 45.0^{o}.

Exercise:

Problem:

Calculate the wavelength of light that has its third minimum at an angle of 30.0^{o} when falling on double slits separated by $3.00\ \mu\text{m}$. Explicitly, show how you follow the steps in Problem-Solving Strategies for Wave Optics.

Solution:

600 nm

Exercise:

Problem:

What is the wavelength of light falling on double slits separated by $2.00\ \mu\text{m}$ if the third-order maximum is at an angle of 60.0^{o}?

Exercise:

Problem:

At what angle is the fourth-order maximum for the situation in [link]?

Solution:

2.06°

Exercise:

Problem:

What is the highest-order maximum for 400-nm light falling on double slits separated by $25.0\ \mu m$?

Exercise:

Problem:

Find the largest wavelength of light falling on double slits separated by $1.20\ \mu m$ for which there is a first-order maximum. Is this in the visible part of the spectrum?

Solution:

1200 nm (not visible)

Exercise:

Problem:

What is the smallest separation between two slits that will produce a second-order maximum for 720-nm red light?

Exercise:

Problem:

(a) What is the smallest separation between two slits that will produce a second-order maximum for any visible light? (b) For all visible light?

Solution:

(a) 760 nm

(b) 1520 nm

Exercise:

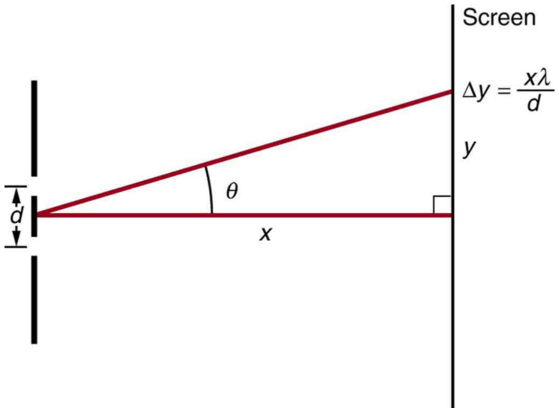
Problem:

(a) If the first-order maximum for pure-wavelength light falling on a double slit is at an angle of 10.0º, at what angle is the second-order maximum? (b) What is the angle of the first minimum? (c) What is the highest-order maximum possible here?

Exercise:

Problem:

[link] shows a double slit located a distance x from a screen, with the distance from the center of the screen given by y. When the distance d between the slits is relatively large, there will be numerous bright spots, called fringes. Show that, for small angles (where $\sin\theta \approx \theta$, with θ in radians), the distance between fringes is given by $\Delta y = x\lambda/d$.



The distance between adjacent fringes is $\Delta y = x\lambda/d$, assuming the slit separation d is large compared with λ.

Solution:

For small angles $\sin\,\theta - \tan\,\theta \approx \theta$ (in radians).

For two adjacent fringes we have,

Equation:

$$d\ \ \sin\,\theta_{\text{m}} = m\lambda$$

and

Equation:

$$d\ \ \sin\,\theta_{\text{m}+1} = (m+1)\lambda$$

Subtracting these equations gives

Equation:

$$\begin{gathered}
d(\sin\,\theta_{\text{m}+1} - \sin\,\theta_{\text{m}}) = [(m+1) - m]\lambda \\
d(\theta_{\text{m}+1}{-}\theta_{\text{m}}) = \lambda \\
\tan\,\theta_{\text{m}} = \frac{y_{\text{m}}}{x} \approx \theta_{\text{m}} \Rightarrow d\ \ \frac{y_{\text{m}+1}}{x} - \frac{y_{\text{m}}}{x}\ \ = \lambda \\
d\frac{\Delta y}{x} = \lambda \Rightarrow \Delta y = \frac{x\lambda}{d}
\end{gathered}$$

Exercise:

Problem:

Using the result of the problem above, calculate the distance between fringes for 633-nm light falling on double slits separated by 0.0800 mm, located 3.00 m from a screen as in [link].

Exercise:

Problem:

Using the result of the problem two problems prior, find the wavelength of light that produces fringes 7.50 mm apart on a screen 2.00 m from double slits separated by 0.120 mm (see [link]).

Solution:

450 nm

Glossary

coherent
: waves are in phase or have a definite phase relationship

constructive interference for a double slit
: the path length difference must be an integral multiple of the wavelength

destructive interference for a double slit
: the path length difference must be a half-integral multiple of the wavelength

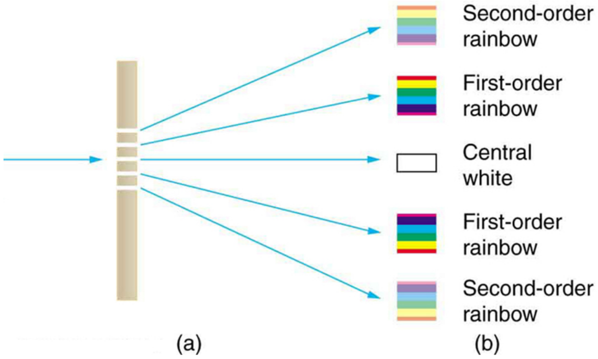
incoherent
: waves have random phase relationships

order
: the integer m used in the equations for constructive and destructive interference for a double slit

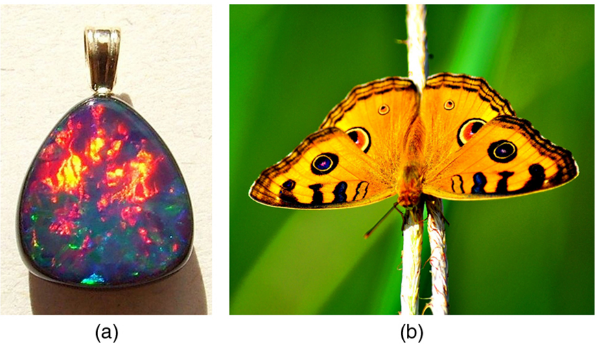
Multiple Slit Diffraction

- Discuss the pattern obtained from diffraction grating.
- Explain diffraction grating effects.

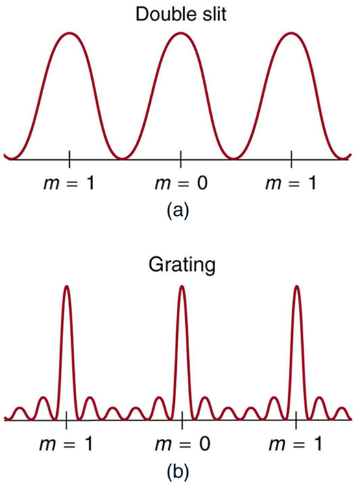
An interesting thing happens if you pass light through a large number of evenly spaced parallel slits, called a **diffraction grating**. An interference pattern is created that is very similar to the one formed by a double slit (see [link]). A diffraction grating can be manufactured by scratching glass with a sharp tool in a number of precisely positioned parallel lines, with the untouched regions acting like slits. These can be photographically mass produced rather cheaply. Diffraction gratings work both for transmission of light, as in [link], and for reflection of light, as on butterfly wings and the Australian opal in [link] or the CD pictured in the opening photograph of this chapter, [link]. In addition to their use as novelty items, diffraction gratings are commonly used for spectroscopic dispersion and analysis of light. What makes them particularly useful is the fact that they form a sharper pattern than double slits do. That is, their bright regions are narrower and brighter, while their dark regions are darker. [link] shows idealized graphs demonstrating the sharper pattern. Natural diffraction gratings occur in the feathers of certain birds. Tiny, finger-like structures in regular patterns act as reflection gratings, producing constructive interference that gives the feathers colors not solely due to their pigmentation. This is called iridescence.



A diffraction grating is a large number of evenly spaced parallel slits. (a) Light passing through is diffracted in a pattern similar to a double slit, with bright regions at various angles. (b) The pattern obtained for white light incident on a grating. The central maximum is white, and the higher-order maxima disperse white light into a rainbow of colors.



(a) This Australian opal and (b) the butterfly wings have rows of reflectors that act like reflection gratings, reflecting different colors at different angles. (credits: (a) Opals-On-Black.com, via Flickr (b) whologwhy, Flickr)



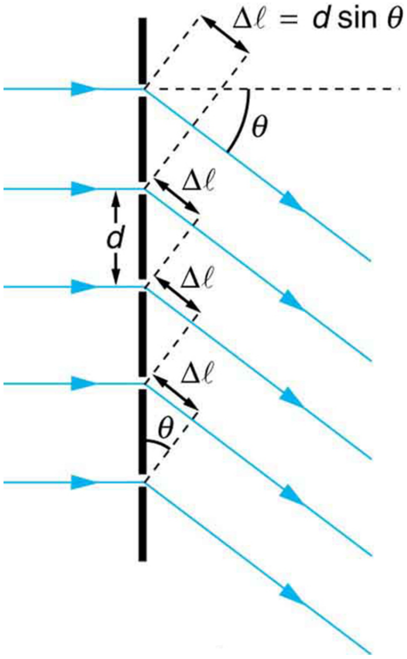
Idealized graphs of the intensity of light passing through a double slit (a) and a diffraction grating (b) for monochromatic light. Maxima can be produced at the same angles, but those for the diffraction grating are narrower and hence sharper. The maxima become narrower and the regions between darker as the number of slits is increased.

The analysis of a diffraction grating is very similar to that for a double slit (see [link]). As we know from our discussion of double slits in Young's Double Slit Experiment, light is diffracted by each slit and spreads out after passing through. Rays traveling in the same direction (at an angle θ relative to the incident direction) are shown in the figure. Each of these rays travels a different distance to a common point on a screen far away. The rays start in phase, and they can be in or out of phase when they reach a screen, depending on the difference in the path lengths traveled. As seen in the figure, each ray travels a distance $d \sin \theta$ different from that of its neighbor, where d is the distance between slits. If this distance equals an integral number of wavelengths, the rays all arrive in phase, and constructive interference (a maximum) is obtained. Thus, the condition necessary to obtain **constructive interference for a diffraction grating** is

Equation:

$$d \ \sin \theta = m\lambda, \text{ for } m = 0, 1, –1, 2, –2, \dots \text{(constructive)},$$

where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum. Note that this is exactly the same equation as for double slits separated by d. However, the slits are usually closer in diffraction gratings than in double slits, producing fewer maxima at larger angles.



Diffraction grating showing light rays from each slit traveling in the same direction. Each ray travels a different distance to reach a common point on a screen (not shown). Each ray travels a distance $d\ \sin\theta$ different from that of its neighbor.

Where are diffraction gratings used? Diffraction gratings are key components of monochromators used, for example, in optical imaging of particular wavelengths from biological or medical samples. A diffraction grating can be chosen to specifically analyze a wavelength emitted by molecules in diseased cells in a biopsy sample or to help excite strategic molecules in the sample with a selected frequency of light. Another vital use is in optical fiber technologies where fibers are designed to provide optimum performance at specific wavelengths. A range of diffraction gratings are available for selecting specific wavelengths for such use.

Note:

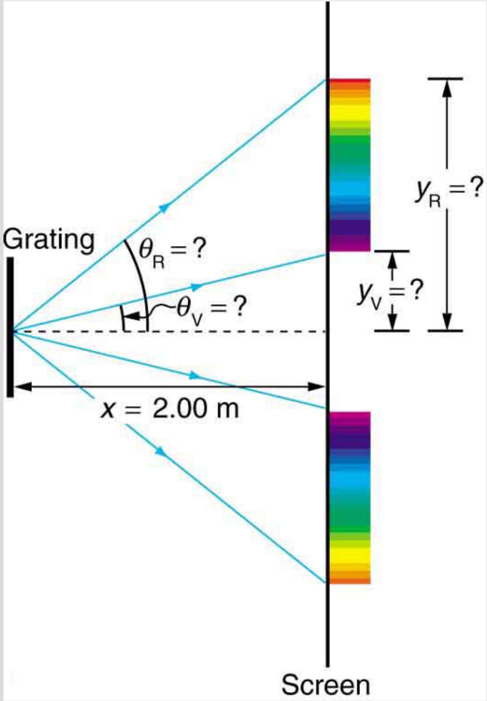
Take-Home Experiment: Rainbows on a CD

The spacing d of the grooves in a CD or DVD can be well determined by using a laser and the equation $d\ \sin\theta = m\lambda,\ \text{for}\ m = 0, 1, –1, 2, –2, \ldots$. However, we can still make a good estimate of this spacing by using white light and the rainbow of colors that comes from the interference. Reflect sunlight from a CD onto a wall and use your best judgment of the location of a strongly diffracted color to find the separation d.

Example:

Calculating Typical Diffraction Grating Effects

Diffraction gratings with 10,000 lines per centimeter are readily available. Suppose you have one, and you send a beam of white light through it to a screen 2.00 m away. (a) Find the angles for the first-order diffraction of the shortest and longest wavelengths of visible light (380 and 760 nm). (b) What is the distance between the ends of the rainbow of visible light produced on the screen for first-order interference? (See [link].)



The diffraction grating considered in this example produces a rainbow of colors on a screen a distance $x = 2.00\text{ m}$ from the grating. The distances along the screen are measured perpendicular to the x-direction. In other words, the rainbow pattern extends out of the page.

Strategy

The angles can be found using the equation

Equation:

$$d\ \sin\theta = m\lambda \text{ (for } m = 0, 1, –1, 2, –2, \ \ldots)$$

once a value for the slit spacing d has been determined. Since there are 10,000 lines per centimeter, each line is separated by $1/10{,}000$ of a centimeter. Once the angles are found, the distances along the screen can be found using simple trigonometry.

Solution for (a)

The distance between slits is $d = (1 \text{ cm})/10{,}000 = 1.00 \times 10^{-4} \text{ cm}$ or 1.00×10^{-6} m. Let us call the two angles θ_{V} for violet (380 nm) and θ_{R} for red (760 nm). Solving the equation d sin $\theta_{\text{V}} = m\lambda$ for sin θ_{V},

Equation:

$$\sin\theta_{\text{V}} = \frac{m\lambda_{\text{V}}}{d},$$

where $m = 1$ for first order and $\lambda_{\text{V}} = 380 \text{ nm} = 3.80 \times 10^{-7}$ m. Substituting these values gives

Equation:

$$\sin\theta_{\text{V}} = \frac{3.80 \times 10^{-7} \text{ m}}{1.00 \times 10^{-6} \text{ m}} = 0.380.$$

Thus the angle θ_{V} is

Equation:

$$\theta_{\text{V}} = \sin^{-1} 0.380 = 22.33^{\text{o}}.$$

Similarly,

Equation:

$$\sin\theta_{\text{R}} = \frac{7.60 \times 10^{-7} \text{ m}}{1.00 \times 10^{-6} \text{ m}}.$$

Thus the angle θ_{R} is

Equation:

$$\theta_{\text{R}} = \sin^{-1} 0.760 = 49.46^{\text{o}}.$$

Notice that in both equations, we reported the results of these intermediate calculations to four significant figures to use with the calculation in part (b).

Solution for (b)

The distances on the screen are labeled y_{V} and y_{R} in [link]. Noting that $\tan\,\theta = y/x$, we can solve for y_{V} and y_{R}. That is,

Equation:

$$y_{\text{V}} = x\,\tan\,\theta_{\text{V}} = (2.00\text{ m})(\tan\,22.33^{\text{o}}) = 0.815\text{ m}$$

and

Equation:

$$y_{\text{R}} = x\,\tan\,\theta_{\text{R}} = (2.00\text{ m})(\tan\,49.46^{\text{o}}) = 2.338\text{ m}.$$

The distance between them is therefore

Equation:

$$y_{\text{R}} - y_{\text{V}} = 1.52\text{ m}.$$

Discussion

The large distance between the red and violet ends of the rainbow produced from the white light indicates the potential this diffraction grating has as a spectroscopic tool. The more it can spread out the wavelengths (greater dispersion), the more detail can be seen in a spectrum. This depends on the quality of the diffraction grating—it must be very precisely made in addition to having closely spaced lines.

Section Summary

- A diffraction grating is a large collection of evenly spaced parallel slits that produces an interference pattern similar to but sharper than that of a double slit.
- There is constructive interference for a diffraction grating when $d\ \sin\theta = m\lambda\ (\text{for}\ m = 0, 1, –1, 2, –2,\ \ldots)$, where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum.

Conceptual Questions

Exercise:

Problem:

What is the advantage of a diffraction grating over a double slit in dispersing light into a spectrum?

Exercise:

Problem:

What are the advantages of a diffraction grating over a prism in dispersing light for spectral analysis?

Exercise:

Problem:

Can the lines in a diffraction grating be too close together to be useful as a spectroscopic tool for visible light? If so, what type of EM radiation would the grating be suitable for? Explain.

Exercise:

Problem:

If a beam of white light passes through a diffraction grating with vertical lines, the light is dispersed into rainbow colors on the right and left. If a glass prism disperses white light to the right into a rainbow, how does the sequence of colors compare with that produced on the right by a diffraction grating?

Exercise:

Problem:

Suppose pure-wavelength light falls on a diffraction grating. What happens to the interference pattern if the same light falls on a grating that has more lines per centimeter? What happens to the interference pattern if a longer-wavelength light falls on the same grating? Explain how these two effects are consistent in terms of the relationship of wavelength to the distance between slits.

Exercise:

Problem:

Suppose a feather appears green but has no green pigment. Explain in terms of diffraction.

Exercise:

Problem:

It is possible that there is no minimum in the interference pattern of a single slit. Explain why. Is the same true of double slits and diffraction gratings?

Problems & Exercises

Exercise:

Problem:

A diffraction grating has 2000 lines per centimeter. At what angle will the first-order maximum be for 520-nm-wavelength green light?

Solution:

5.97^{o}

Exercise:

Problem:

Find the angle for the third-order maximum for 580-nm-wavelength yellow light falling on a diffraction grating having 1500 lines per centimeter.

Exercise:

Problem:

How many lines per centimeter are there on a diffraction grating that gives a first-order maximum for 470-nm blue light at an angle of 25.0° ?

Solution:

8.99×10^{3}

Exercise:

Problem:

What is the distance between lines on a diffraction grating that produces a second-order maximum for 760-nm red light at an angle of 60.0°?

Exercise:

Problem:

Calculate the wavelength of light that has its second-order maximum at 45.0° when falling on a diffraction grating that has 5000 lines per centimeter.

Solution:

707 nm

Exercise:

Problem:

An electric current through hydrogen gas produces several distinct wavelengths of visible light. What are the wavelengths of the hydrogen spectrum, if they form first-order maxima at angles of 24.2°, 25.7°, 29.1°, and 41.0° when projected on a diffraction grating having 10,000 lines per centimeter? Explicitly show how you follow the steps in Problem-Solving Strategies for Wave Optics

Exercise:

Problem:

(a) What do the four angles in the above problem become if a 5000-line-per-centimeter diffraction grating is used? (b) Using this grating, what would the angles be for the second-order maxima? (c) Discuss the relationship between integral reductions in lines per centimeter and the new angles of various order maxima.

Solution:

$\text{(a)}\ \ 11.8^{\circ},\ 12.5^{\circ},\ 14.1^{\circ},\ 19.2^{\circ}$

$\text{(b)}\ \ 24.2^{\circ},\ 25.7^{\circ},\ 29.1^{\circ},\ 41.0^{\circ}$

(c) Decreasing the number of lines per centimeter by a factor of x means that the angle for the x-order maximum is the same as the original angle for the first- order maximum.

Exercise:

Problem:

What is the maximum number of lines per centimeter a diffraction grating can have and produce a complete first-order spectrum for visible light?

Exercise:

Problem:

The yellow light from a sodium vapor lamp *seems* to be of pure wavelength, but it produces two first-order maxima at 36.093° and 36.129° when projected on a 10,000 line per centimeter diffraction grating. What are the two wavelengths to an accuracy of 0.1 nm?

Solution:

589.1 nm and 589.6 nm

Exercise:

Problem:

What is the spacing between structures in a feather that acts as a reflection grating, given that they produce a first-order maximum for 525-nm light at a 30.0° angle?

Exercise:

Problem:

Structures on a bird feather act like a reflection grating having 8000 lines per centimeter. What is the angle of the first-order maximum for 600-nm light?

Solution:

28.7°

Exercise:

Problem:

An opal such as that shown in [link] acts like a reflection grating with rows separated by about $8\ \mu m$. If the opal is illuminated normally, (a) at what angle will red light be seen and (b) at what angle will blue light be seen?

Exercise:

Problem:

At what angle does a diffraction grating produces a second-order maximum for light having a first-order maximum at 20.0°?

Solution:

43.2°

Exercise:

Problem:

Show that a diffraction grating cannot produce a second-order maximum for a given wavelength of light unless the first-order maximum is at an angle less than 30.0°.

Exercise:

Problem:

If a diffraction grating produces a first-order maximum for the shortest wavelength of visible light at 30.0°, at what angle will the first-order maximum be for the longest wavelength of visible light?

Solution:

90.0°

Exercise:

Problem:

(a) Find the maximum number of lines per centimeter a diffraction grating can have and produce a maximum for the smallest wavelength of visible light. (b) Would such a grating be useful for ultraviolet spectra? (c) For infrared spectra?

Exercise:

Problem:

(a) Show that a 30,000-line-per-centimeter grating will not produce a maximum for visible light. (b) What is the longest wavelength for which it does produce a first-order maximum? (c) What is the greatest number of lines per centimeter a diffraction grating can have and produce a complete second-order spectrum for visible light?

Solution:

(a) The longest wavelength is 333.3 nm, which is not visible.

(b) 333 nm (UV)

(c) 6.58×10^3 cm

Exercise:

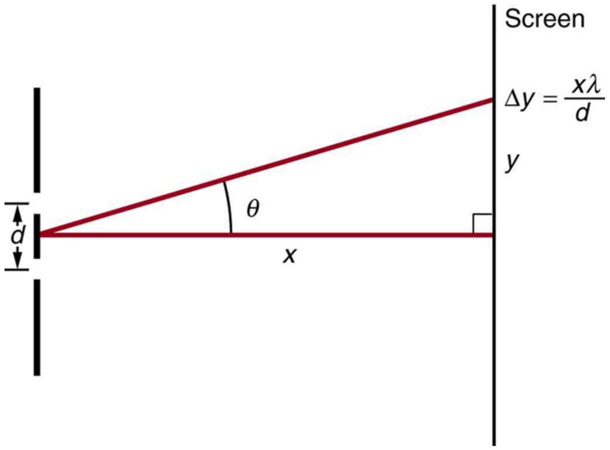
Problem:

A He–Ne laser beam is reflected from the surface of a CD onto a wall. The brightest spot is the reflected beam at an angle equal to the angle of incidence. However, fringes are also observed. If the wall is 1.50 m from the CD, and the first fringe is 0.600 m from the central maximum, what is the spacing of grooves on the CD?

Exercise:

Problem:

The analysis shown in the figure below also applies to diffraction gratings with lines separated by a distance d. What is the distance between fringes produced by a diffraction grating having 125 lines per centimeter for 600-nm light, if the screen is 1.50 m away?



The distance between adjacent fringes is $\Delta y = x\lambda/d$, assuming the slit separation d is large compared with λ.

Solution:

$1.13 \times 10^{-2}\ \text{m}$

Exercise:

Problem: Unreasonable Results

Red light of wavelength of 700 nm falls on a double slit separated by 400 nm. (a) At what angle is the first-order maximum in the diffraction pattern? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: Unreasonable Results

(a) What visible wavelength has its fourth-order maximum at an angle of 25.0° when projected on a 25,000-line-per-centimeter diffraction grating? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) 42.3 nm

(b) Not a visible wavelength

The number of slits in this diffraction grating is too large. Etching in integrated circuits can be done to a resolution of 50 nm, so slit separations of 400 nm are at the limit of what we can do today. This line spacing is too small to produce diffraction of light.

Exercise:

Problem: Construct Your Own Problem

Consider a spectrometer based on a diffraction grating. Construct a problem in which you calculate the distance between two wavelengths of electromagnetic radiation in your spectrometer. Among the things to be considered are the wavelengths you wish to be able to distinguish, the number of lines per meter on the diffraction grating, and the distance from the grating to the screen or detector. Discuss the practicality of the device in terms of being able to discern between wavelengths of interest.

Glossary

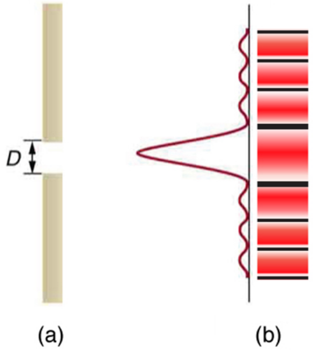
constructive interference for a diffraction grating
: occurs when the condition $d\ \sin\theta = m\lambda\ (\text{for}\ m = 0,\, 1,\, –1,\, 2,\, –2,\ \ldots)$ is satisfied, where d is the distance between slits in the grating, λ is the wavelength of light, and m is the order of the maximum

diffraction grating
: a large number of evenly spaced parallel slits

Single Slit Diffraction

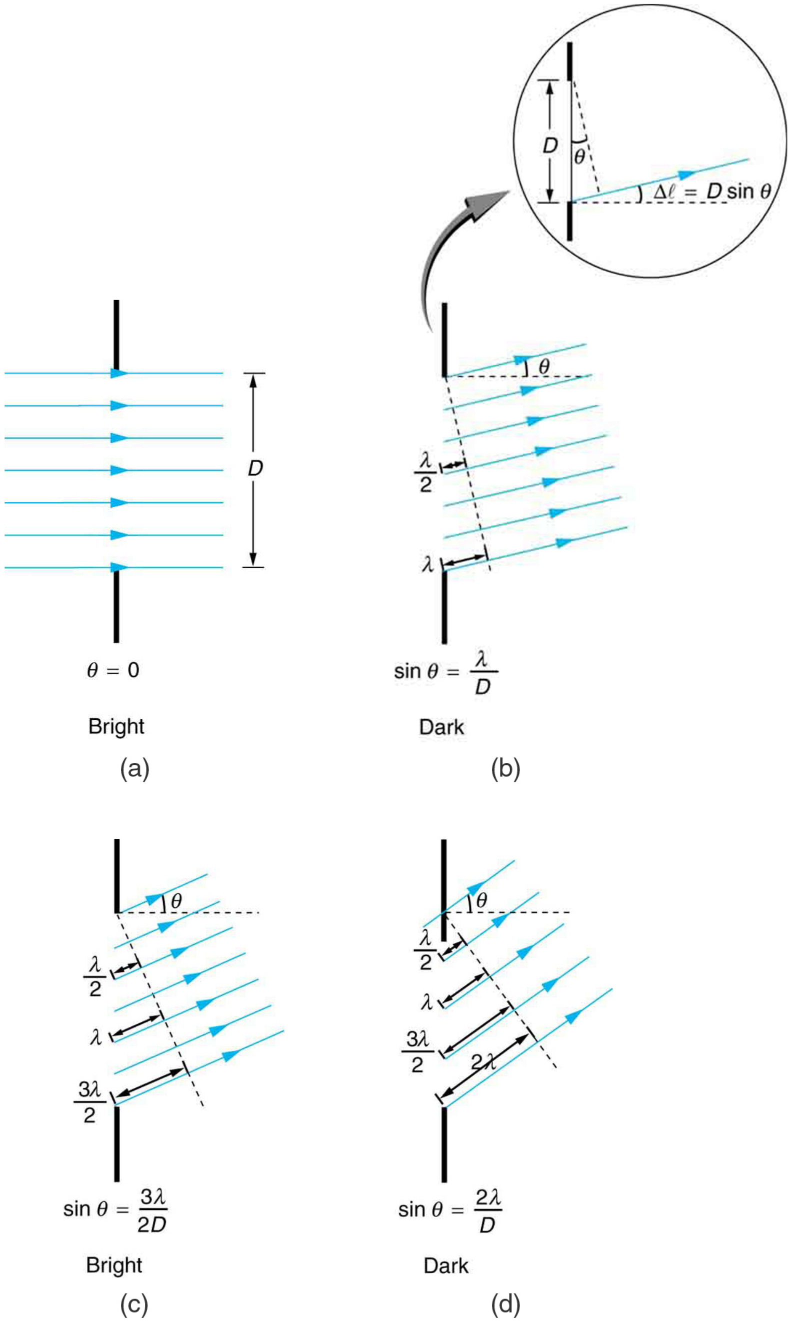
- Discuss the single slit diffraction pattern.

Light passing through a single slit forms a diffraction pattern somewhat different from those formed by double slits or diffraction gratings. [link] shows a single slit diffraction pattern. Note that the central maximum is larger than those on either side, and that the intensity decreases rapidly on either side. In contrast, a diffraction grating produces evenly spaced lines that dim slowly on either side of center.



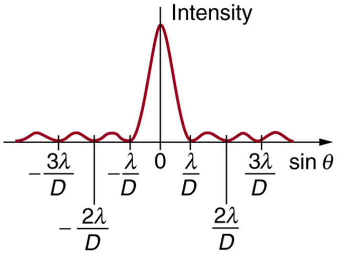
(a) Single slit diffraction pattern. Monochromatic light passing through a single slit has a central maximum and many smaller and dimmer maxima on either side. The central maximum is six times higher than shown. (b) The drawing shows the bright central maximum and dimmer and thinner maxima on either side.

The analysis of single slit diffraction is illustrated in [link]. Here we consider light coming from different parts of the *same* slit. According to Huygens’s principle, every part of the wavefront in the slit emits wavelets. These are like rays that start out in phase and head in all directions. (Each ray is perpendicular to the wavefront of a wavelet.) Assuming the screen is very far away compared with the size of the slit, rays heading toward a common destination are nearly parallel. When they travel straight ahead, as in [link](a), they remain in phase, and a central maximum is obtained. However, when rays travel at an angle θ relative to the original direction of the beam, each travels a different distance to a common location, and they can arrive in or out of phase. In [link](b), the ray from the bottom travels a distance of one wavelength λ farther than the ray from the top. Thus a ray from the center travels a distance $\lambda/2$ farther than the one on the left, arrives out of phase, and interferes destructively. A ray from slightly above the center and one from slightly above the bottom will also cancel one another. In fact, each ray from the slit will have another to interfere destructively, and a minimum in intensity will occur at this angle. There will be another minimum at the same angle to the right of the incident direction of the light.



Light passing through a single slit is diffracted in all directions and may interfere constructively or destructively, depending on the angle. The difference in path length for rays from either side of the slit is seen to be $D \sin \theta$.

At the larger angle shown in [link](c), the path lengths differ by $3\lambda/2$ for rays from the top and bottom of the slit. One ray travels a distance λ different from the ray from the bottom and arrives in phase, interfering constructively. Two rays, each from slightly above those two, will also add constructively. Most rays from the slit will have another to interfere with constructively, and a maximum in intensity will occur at this angle. However, all rays do not interfere constructively for this situation, and so the maximum is not as intense as the central maximum. Finally, in [link](d), the angle shown is large enough to produce a second minimum. As seen in the figure, the difference in path length for rays from either side of the slit is $D\ \sin\ \theta$, and we see that a destructive minimum is obtained when this distance is an integral multiple of the wavelength.



A graph of single slit diffraction intensity showing the central maximum to be wider and much more intense than those to the sides. In fact the central maximum is six times higher than shown here.

Thus, to obtain **destructive interference for a single slit**,

Equation:

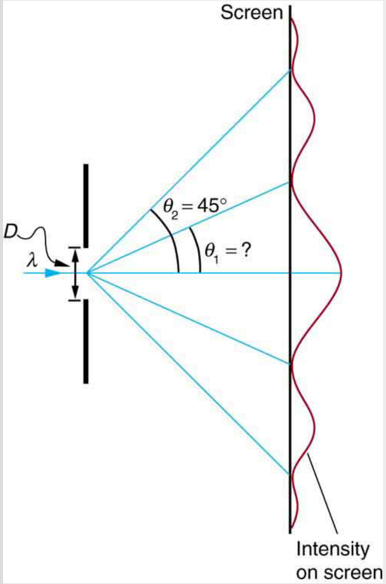
$$D\sin\theta = m\lambda, \text{ for } m = 1, –1, 2, –2, 3, \ \ldots \ \text{(destructive)},$$

where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. [link] shows a graph of intensity for single slit interference, and it is apparent that the maxima on either side of the central maximum are much less intense and not as wide. This is consistent with the illustration in [link] (b).

Example:

Calculating Single Slit Diffraction

Visible light of wavelength 550 nm falls on a single slit and produces its second diffraction minimum at an angle of 45.0º relative to the incident direction of the light. (a) What is the width of the slit? (b) At what angle is the first minimum produced?



A graph of the

single slit diffraction pattern is analyzed in this example.

Strategy

From the given information, and assuming the screen is far away from the slit, we can use the equation $D \sin \theta = \mathrm{m}\lambda$ first to find D, and again to find the angle for the first minimum θ_1.

Solution for (a)

We are given that $\lambda = 550 \text{ nm}$, $m = 2$, and $\theta_2 = 45.0^\circ$. Solving the equation $D \ \sin \theta = \mathrm{m}\lambda$ for D and substituting known values gives

Equation:

$$\begin{aligned} D &= \frac{m\lambda}{\sin \theta_2} = \frac{2(550 \text{ nm})}{\sin 45.0^\circ} \\ &= \frac{1100 \times 10^{-9}}{0.707} \\ &= 1.56 \times 10^{-6}. \end{aligned}$$

Solution for (b)

Solving the equation $D \ \sin \theta = m\lambda$ for $\sin \theta_1$ and substituting the known values gives

Equation:

$$\sin \theta_1 = \frac{m\lambda}{D} = \frac{1 \ \ 550 \times 10^{-9} \text{ m}}{1.56 \times 10^{-6} \text{ m}}.$$

Thus the angle θ_1 is

Equation:

$$\theta_1 = \sin^{-1} 0.354 = 20.7^\circ.$$

Discussion

We see that the slit is narrow (it is only a few times greater than the wavelength of light). This is consistent with the fact that light must interact with an object comparable in size to its wavelength in order to exhibit significant wave effects such as this single slit diffraction pattern. We also see that the central maximum extends 20.7^{o} on either side of the original beam, for a width of about 41^{o} . The angle between the first and second minima is only about 24^{o} $(45.0^{\mathrm{o}} - 20.7^{\mathrm{o}})$. Thus the second maximum is only about half as wide as the central maximum.

Section Summary

- A single slit produces an interference pattern characterized by a broad central maximum with narrower and dimmer maxima to the sides.
- There is destructive interference for a single slit when $D\ \sin\theta = m\lambda,$ (for $m = 1$, –1, 2, –2, 3, $\ldots$), where D is the slit width, λ is the light's wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum. Note that there is no $m = 0$ minimum.

Conceptual Questions

Exercise:

Problem:

As the width of the slit producing a single-slit diffraction pattern is reduced, how will the diffraction pattern produced change?

Problems & Exercises

Exercise:

Problem:

(a) At what angle is the first minimum for 550-nm light falling on a single slit of width 1.00 μm? (b) Will there be a second minimum?

Solution:

(a) 33.4°

(b) No

Exercise:

Problem:

(a) Calculate the angle at which a $2.00\text{-}\mu\text{m}$-wide slit produces its first minimum for 410-nm violet light. (b) Where is the first minimum for 700-nm red light?

Exercise:

Problem:

(a) How wide is a single slit that produces its first minimum for 633-nm light at an angle of 28.0°? (b) At what angle will the second minimum be?

Solution:

(a) $1.35 \times 10^{-6}\text{ m}$

(b) 69.9°

Exercise:

Problem:

(a) What is the width of a single slit that produces its first minimum at 60.0° for 600-nm light? (b) Find the wavelength of light that has its first minimum at 62.0°.

Exercise:

Problem:

Find the wavelength of light that has its third minimum at an angle of 48.6° when it falls on a single slit of width $3.00\ \mu\text{m}$.

Solution:

750 nm

Exercise:

Problem:

Calculate the wavelength of light that produces its first minimum at an angle of 36.9° when falling on a single slit of width $1.00\ \mu m$.

Exercise:

Problem:

(a) Sodium vapor light averaging 589 nm in wavelength falls on a single slit of width $7.50\ \mu m$. At what angle does it produces its second minimum? (b) What is the highest-order minimum produced?

Solution:

(a) 9.04°

(b) 12

Exercise:

Problem:

(a) Find the angle of the third diffraction minimum for 633-nm light falling on a slit of width $20.0\ \mu m$. (b) What slit width would place this minimum at 85.0°? Explicitly show how you follow the steps in Problem-Solving Strategies for Wave Optics

Exercise:

Problem:

(a) Find the angle between the first minima for the two sodium vapor lines, which have wavelengths of 589.1 and 589.6 nm, when they fall upon a single slit of width $2.00\ \mu m$. (b) What is the distance between these minima if the diffraction pattern falls on a screen 1.00 m from the slit? (c) Discuss the ease or difficulty of measuring such a distance.

Solution:

(a) 0.0150^o

(b) 0.262 mm

(c) This distance is not easily measured by human eye, but under a microscope or magnifying glass it is quite easily measurable.

Exercise:

Problem:

(a) What is the minimum width of a single slit (in multiples of λ) that will produce a first minimum for a wavelength λ? (b) What is its minimum width if it produces 50 minima? (c) 1000 minima?

Exercise:

Problem:

(a) If a single slit produces a first minimum at 14.5^o, at what angle is the second-order minimum? (b) What is the angle of the third-order minimum? (c) Is there a fourth-order minimum? (d) Use your answers to illustrate how the angular width of the central maximum is about twice the angular width of the next maximum (which is the angle between the first and second minima).

Solution:

(a) 30.1^o

(b) 48.7^o

(c) No

(d) $2\text{θ}_1 = (2)(14.5^o) = 29^o$, $\theta_2 - \theta_1 = 30.05^o - 14.5^o$=$15.56^o$. Thus, $29^o \approx (2)(15.56^o) = 31.1^o$.

Exercise:

Problem:

A double slit produces a diffraction pattern that is a combination of single and double slit interference. Find the ratio of the width of the slits to the separation between them, if the first minimum of the single slit pattern falls on the fifth maximum of the double slit pattern. (This will greatly reduce the intensity of the fifth maximum.)

Exercise:

Problem: Integrated Concepts

A water break at the entrance to a harbor consists of a rock barrier with a 50.0-m-wide opening. Ocean waves of 20.0-m wavelength approach the opening straight on. At what angle to the incident direction are the boats inside the harbor most protected against wave action?

Solution:

23.6° and 53.1°

Exercise:

Problem: Integrated Concepts

An aircraft maintenance technician walks past a tall hangar door that acts like a single slit for sound entering the hangar. Outside the door, on a line perpendicular to the opening in the door, a jet engine makes a 600-Hz sound. At what angle with the door will the technician observe the first minimum in sound intensity if the vertical opening is 0.800 m wide and the speed of sound is 340 m/s?

Glossary

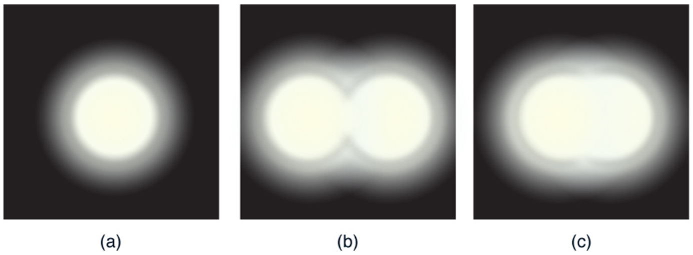
destructive interference for a single slit

occurs when $D \sin\theta = \mathrm{m}\lambda$, (for $m = 1, –1, 2, –2, 3, \ldots$), where D is the slit width, λ is the light’s wavelength, θ is the angle relative to the original direction of the light, and m is the order of the minimum

Limits of Resolution: The Rayleigh Criterion

- Discuss the Rayleigh criterion.

Light diffracts as it moves through space, bending around obstacles, interfering constructively and destructively. While this can be used as a spectroscopic tool—a diffraction grating disperses light according to wavelength, for example, and is used to produce spectra—diffraction also limits the detail we can obtain in images. [link](a) shows the effect of passing light through a small circular aperture. Instead of a bright spot with sharp edges, a spot with a fuzzy edge surrounded by circles of light is obtained. This pattern is caused by diffraction similar to that produced by a single slit. Light from different parts of the circular aperture interferes constructively and destructively. The effect is most noticeable when the aperture is small, but the effect is there for large apertures, too.



(a) Monochromatic light passed through a small circular aperture produces this diffraction pattern. (b) Two point light sources that are close to one another produce overlapping images because of diffraction. (c) If they are closer together, they cannot be resolved or distinguished.

How does diffraction affect the detail that can be observed when light passes through an aperture? [link](b) shows the diffraction pattern produced by two point light sources that are close to one another. The pattern is similar to that for a single point source, and it is just barely possible to tell that there are two light sources rather than one. If they were closer together, as in [link](c), we could not distinguish them, thus limiting the detail or resolution we can obtain. This limit is an inescapable consequence of the wave nature of light.

There are many situations in which diffraction limits the resolution. The acuity of our vision is limited because light passes through the pupil, the circular aperture of our eye. Be aware that the diffraction-like spreading of light is due to the limited diameter of a light beam, not the interaction with an aperture. Thus light passing through a lens with a diameter D shows this effect and spreads, blurring the image, just as light passing through an aperture of diameter D does. So diffraction limits the resolution of any system having a lens or mirror. Telescopes are also limited by diffraction, because of the finite diameter D of their primary mirror.

Note:

Take-Home Experiment: Resolution of the Eye

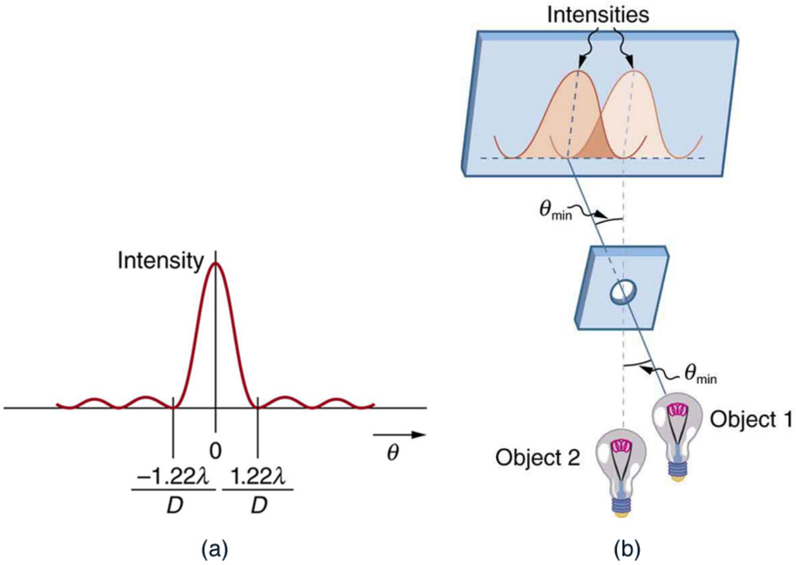
Draw two lines on a white sheet of paper (several mm apart). How far away can you be and still distinguish the two lines? What does this tell you about the size of the eye's pupil? Can you be quantitative? (The size of an adult's pupil is discussed in Physics of the Eye.)

Just what is the limit? To answer that question, consider the diffraction pattern for a circular aperture, which has a central maximum that is wider and brighter than the maxima surrounding it (similar to a slit) [see [link] (a)]. It can be shown that, for a circular aperture of diameter D, the first minimum in the diffraction pattern occurs at $\theta = 1.22\ \lambda/D$ (providing the aperture is large compared with the wavelength of light, which is the case for most optical instruments). The accepted criterion for determining the diffraction limit to resolution based on this angle was developed by Lord Rayleigh in the 19th century. The **Rayleigh criterion** for the diffraction limit to resolution states that *two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other*. See [link](b). The first minimum is at an angle of $\theta = 1.22\ \lambda/D$, so that two point objects are just resolvable if they are separated by the angle

Equation:

$$\theta = 1.22\frac{\lambda}{D},$$

where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc., with which the two objects are observed. In this expression, θ has units of radians.



(a) Graph of intensity of the diffraction pattern for a circular aperture. Note that, similar to a single slit, the central maximum is wider and brighter than those to the sides. (b) Two point objects produce overlapping diffraction patterns. Shown here is the Rayleigh criterion for being just resolvable. The central maximum of one pattern lies on the first minimum of the other.

Note:

Connections: Limits to Knowledge

All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Even the small wavelength of light prohibits exact precision. When extremely small wavelength probes as with an electron microscope are used, the system is disturbed, still limiting our knowledge, much as making an electrical measurement alters a circuit. Heisenberg’s uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

Example:

Calculating Diffraction Limits of the Hubble Space Telescope

The primary mirror of the orbiting Hubble Space Telescope has a diameter of 2.40 m. Being in orbit, this telescope avoids the degrading effects of atmospheric distortion on its resolution. (a) What is the angle between two just-resolvable point light sources (perhaps two stars)? Assume an average light wavelength of 550 nm. (b) If these two stars are at the 2 million light year distance of the Andromeda galaxy, how close together can they be and still be resolved? (A light year, or ly, is the distance light travels in 1 year.)

Strategy

The Rayleigh criterion stated in the equation $\theta = 1.22\frac{\lambda}{D}$ gives the smallest possible angle θ between point sources, or the best obtainable resolution. Once this angle is found, the distance between stars can be calculated, since we are given how far away they are.

Solution for (a)

The Rayleigh criterion for the minimum resolvable angle is

Equation:

$$\theta = 1.22\frac{\lambda}{D}.$$

Entering known values gives

Equation:

$$\begin{aligned}\theta &= 1.22\,\frac{550\times10^{-9}\text{ m}}{2.40\text{ m}}\\ &= 2.80\times10^{-7}\text{ rad.}\end{aligned}$$

Solution for (b)

The distance s between two objects a distance r away and separated by an angle θ is $s=r\theta$.

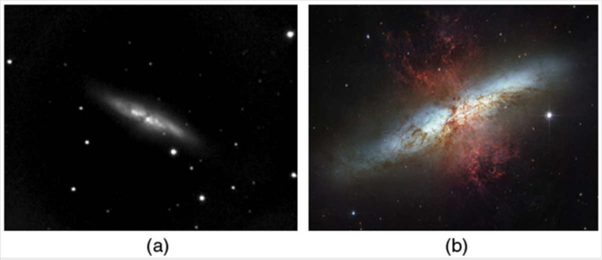
Substituting known values gives

Equation:

$$\begin{aligned}s &= (2.0\times10^{6}\text{ ly})(2.80\times10^{-7}\text{ rad})\\ &= 0.56\text{ ly.}\end{aligned}$$

Discussion

The angle found in part (a) is extraordinarily small (less than 1/50,000 of a degree), because the primary mirror is so large compared with the wavelength of light. As noticed, diffraction effects are most noticeable when light interacts with objects having sizes on the order of the wavelength of light. However, the effect is still there, and there is a diffraction limit to what is observable. The actual resolution of the Hubble Telescope is not quite as good as that found here. As with all instruments, there are other effects, such as non-uniformities in mirrors or aberrations in lenses that further limit resolution. However, [link] gives an indication of the extent of the detail observable with the Hubble because of its size and quality and especially because it is above the Earth’s atmosphere.



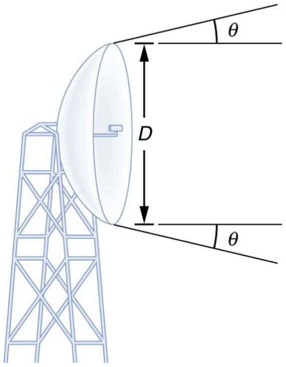
These two photographs of the M82 galaxy give an idea of the observable detail using the Hubble Space Telescope compared with that using a ground-based telescope. (a) On the left is a ground-based image. (credit: Ricnun, Wikimedia Commons) (b) The photo on the right was captured by Hubble. (credit: NASA, ESA, and the Hubble Heritage Team (STScI/AURA))

The answer in part (b) indicates that two stars separated by about half a light year can be resolved. The average distance between stars in a galaxy is on the order of 5 light years in the outer parts and about 1 light year near the galactic center. Therefore, the Hubble can resolve most of the individual stars in Andromeda galaxy, even though it lies at such a huge distance that its light takes 2 million years for its light to reach us. [link] shows another mirror used to observe radio waves from outer space.

A 305-m-diameter natural bowl at Arecibo in Puerto Rico is lined with reflective material, making it into a radio telescope. It is the largest curved focusing dish in the world. Although D for Arecibo is much larger than for the

Hubble Telescope, it detects much longer wavelength radiation and its diffraction limit is significantly poorer than Hubble’s. Arecibo is still very useful, because important information is carried by radio waves that is not carried by visible light. (credit: Tatyana Temirbulatova, Flickr)

Diffraction is not only a problem for optical instruments but also for the electromagnetic radiation itself. Any beam of light having a finite diameter D and a wavelength λ exhibits diffraction spreading. The beam spreads out with an angle θ given by the equation $\theta = 1.22\frac{\lambda}{D}$. Take, for example, a laser beam made of rays as parallel as possible (angles between rays as close to $\theta = 0^\text{o}$ as possible) instead spreads out at an angle $\theta = 1.22\ \lambda/D$, where D is the diameter of the beam and λ is its wavelength. This spreading is impossible to observe for a flashlight, because its beam is not very parallel to start with. However, for long-distance transmission of laser beams or microwave signals, diffraction spreading can be significant (see [link]). To avoid this, we can increase D. This is done for laser light sent to the Moon to measure its distance from the Earth. The laser beam is expanded through a telescope to make D much larger and θ smaller.



The beam produced by this microwave transmission antenna will spread out at a minimum angle $\theta = 1.22\ \lambda/D$ due to diffraction. It is impossible to produce a near-parallel beam, because the beam has a limited diameter.

In most biology laboratories, resolution is presented when the use of the microscope is introduced. The ability of a lens to produce sharp images of two closely spaced point objects is called resolution. The smaller the distance x by which two objects can be separated and still be seen as distinct, the greater the resolution. The resolving power of a lens is defined as that distance x. An expression for resolving power is obtained from the

Rayleigh criterion. In [link](a) we have two point objects separated by a distance x. According to the Rayleigh criterion, resolution is possible when the minimum angular separation is

Equation:

$$\theta = 1.22\frac{\lambda}{D} = \frac{x}{d},$$

where d is the distance between the specimen and the objective lens, and we have used the small angle approximation (i.e., we have assumed that x is much smaller than d), so that $\tan\,\theta \approx \sin\,\theta \approx \theta$.

Therefore, the resolving power is

Equation:

$$x = 1.22\frac{\lambda d}{D}.$$

Another way to look at this is by re-examining the concept of Numerical Aperture (NA) discussed in Microscopes. There, NA is a measure of the maximum acceptance angle at which the fiber will take light and still contain it within the fiber. [link](b) shows a lens and an object at point P. The NA here is a measure of the ability of the lens to gather light and resolve fine detail. The angle subtended by the lens at its focus is defined to be $\theta = 2\alpha$. From the figure and again using the small angle approximation, we can write

Equation:

$$\sin\,\alpha = \frac{D/2}{d} = \frac{D}{2d}.$$

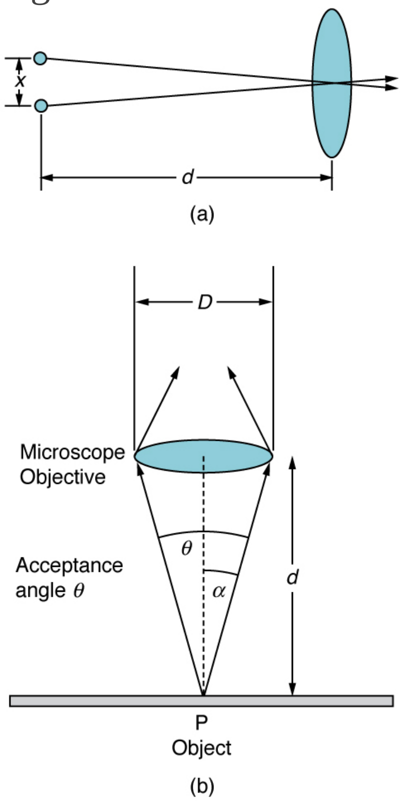
The NA for a lens is $NA = n\,\sin\,\alpha$, where n is the index of refraction of the medium between the objective lens and the object at point P.

From this definition for NA, we can see that

Equation:

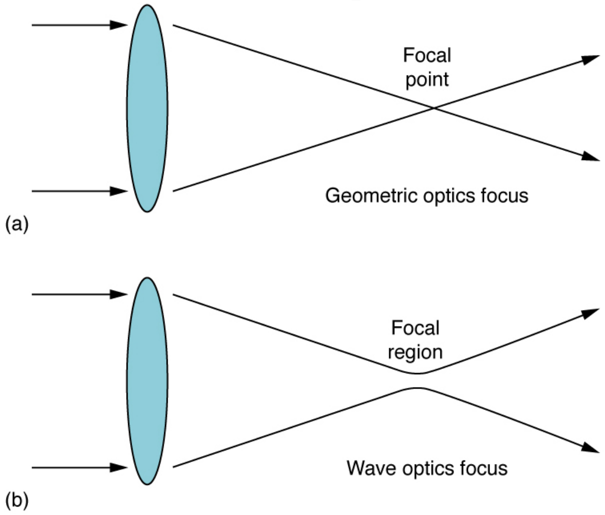
$$x = 1.22\frac{\lambda d}{D} = 1.22\frac{\lambda}{2\ \sin\alpha} = 0.61\frac{\lambda n}{\text{NA}}.$$

In a microscope, NA is important because it relates to the resolving power of a lens. A lens with a large NA will be able to resolve finer details. Lenses with larger NA will also be able to collect more light and so give a brighter image. Another way to describe this situation is that the larger the NA, the larger the cone of light that can be brought into the lens, and so more of the diffraction modes will be collected. Thus the microscope has more information to form a clear image, and so its resolving power will be higher.



(a) Two points separated by at distance x and a positioned a distance d away from the objective. (credit: Infopro, Wikimedia Commons) (b) Terms and symbols used in discussion of resolving power for a lens and an object at point P. (credit: Infopro, Wikimedia Commons)

One of the consequences of diffraction is that the focal point of a beam has a finite width and intensity distribution. Consider focusing when only considering geometric optics, shown in [link](a). The focal point is infinitely small with a huge intensity and the capacity to incinerate most samples irrespective of the NA of the objective lens. For wave optics, due to diffraction, the focal point spreads to become a focal spot (see [link](b)) with the size of the spot decreasing with increasing NA. Consequently, the intensity in the focal spot increases with increasing NA. The higher the NA , the greater the chances of photodegrading the specimen. However, the spot never becomes a true point.



(a) In geometric optics, the focus is a point, but it is not physically possible to produce such a point because it implies infinite intensity. (b) In wave optics, the focus is an extended region.

Section Summary

- Diffraction limits resolution.
- For a circular aperture, lens, or mirror, the Rayleigh criterion states that two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other.
- This occurs for two point objects separated by the angle $\theta = 1.22\frac{\lambda}{D}$, where λ is the wavelength of light (or other electromagnetic radiation) and D is the diameter of the aperture, lens, mirror, etc. This equation also gives the angular spreading of a source of light having a diameter D.

Conceptual Questions

Exercise:

Problem:

A beam of light always spreads out. Why can a beam not be created with parallel rays to prevent spreading? Why can lenses, mirrors, or apertures not be used to correct the spreading?

Problems & Exercises

Exercise:

Problem:

The 300-m-diameter Arecibo radio telescope pictured in [link] detects radio waves with a 4.00 cm average wavelength.

(a) What is the angle between two just-resolvable point sources for this telescope?

(b) How close together could these point sources be at the 2 million light year distance of the Andromeda galaxy?

Solution:

(a) $1.63 \times 10^{-4}\ \text{rad}$

(b) 326 ly

Exercise:

Problem:

Assuming the angular resolution found for the Hubble Telescope in [link], what is the smallest detail that could be observed on the Moon?

Exercise:

Problem:

Diffraction spreading for a flashlight is insignificant compared with other limitations in its optics, such as spherical aberrations in its mirror. To show this, calculate the minimum angular spreading of a flashlight beam that is originally 5.00 cm in diameter with an average wavelength of 600 nm.

Solution:

$1.46 \times 10^{-5}\ \text{rad}$

Exercise:

Problem:

(a) What is the minimum angular spread of a 633-nm wavelength He-Ne laser beam that is originally 1.00 mm in diameter?

(b) If this laser is aimed at a mountain cliff 15.0 km away, how big will the illuminated spot be?

(c) How big a spot would be illuminated on the Moon, neglecting atmospheric effects? (This might be done to hit a corner reflector to measure the round-trip time and, hence, distance.) Explicitly show how you follow the steps in Problem-Solving Strategies for Wave Optics.

Exercise:

Problem:

A telescope can be used to enlarge the diameter of a laser beam and limit diffraction spreading. The laser beam is sent through the telescope in opposite the normal direction and can then be projected onto a satellite or the Moon.

(a) If this is done with the Mount Wilson telescope, producing a 2.54-m-diameter beam of 633-nm light, what is the minimum angular spread of the beam?

(b) Neglecting atmospheric effects, what is the size of the spot this beam would make on the Moon, assuming a lunar distance of 3.84×10^8 m?

Solution:

(a) 3.04×10^{-7} rad

(b) Diameter of 235 m

Exercise:

Problem:

The limit to the eye’s acuity is actually related to diffraction by the pupil.

(a) What is the angle between two just-resolvable points of light for a 3.00-mm-diameter pupil, assuming an average wavelength of 550 nm?

(b) Take your result to be the practical limit for the eye. What is the greatest possible distance a car can be from you if you can resolve its two headlights, given they are 1.30 m apart?

(c) What is the distance between two just-resolvable points held at an arm’s length (0.800 m) from your eye?

(d) How does your answer to (c) compare to details you normally observe in everyday circumstances?

Exercise:

Problem:

What is the minimum diameter mirror on a telescope that would allow you to see details as small as 5.00 km on the Moon some 384,000 km away? Assume an average wavelength of 550 nm for the light received.

Solution:

5.15 cm

Exercise:

Problem:

You are told not to shoot until you see the whites of their eyes. If the eyes are separated by 6.5 cm and the diameter of your pupil is 5.0 mm, at what distance can you resolve the two eyes using light of wavelength 555 nm?

Exercise:

Problem:

(a) The planet Pluto and its Moon Charon are separated by 19,600 km. Neglecting atmospheric effects, should the 5.08-m-diameter Mount Palomar telescope be able to resolve these bodies when they are 4.50×10^9 km from Earth? Assume an average wavelength of 550 nm.

(b) In actuality, it is just barely possible to discern that Pluto and Charon are separate bodies using an Earth-based telescope. What are the reasons for this?

Solution:

(a) Yes. Should easily be able to discern.

(b) The fact that it is just barely possible to discern that these are separate bodies indicates the severity of atmospheric aberrations.

Exercise:

Problem:

The headlights of a car are 1.3 m apart. What is the maximum distance at which the eye can resolve these two headlights? Take the pupil diameter to be 0.40 cm.

Exercise:

Problem:

When dots are placed on a page from a laser printer, they must be close enough so that you do not see the individual dots of ink. To do this, the separation of the dots must be less than Raleigh's criterion. Take the pupil of the eye to be 3.0 mm and the distance from the paper to the eye of 35 cm; find the minimum separation of two dots such that they cannot be resolved. How many dots per inch (dpi) does this correspond to?

Exercise:

Problem: Unreasonable Results

An amateur astronomer wants to build a telescope with a diffraction limit that will allow him to see if there are people on the moons of Jupiter.

(a) What diameter mirror is needed to be able to see 1.00 m detail on a Jovian Moon at a distance of $7.50 \times 10^8 \text{ km}$ from Earth? The wavelength of light averages 600 nm.

(b) What is unreasonable about this result?

(c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: Construct Your Own Problem

Consider diffraction limits for an electromagnetic wave interacting with a circular object. Construct a problem in which you calculate the limit of angular resolution with a device, using this circular object (such as a lens, mirror, or antenna) to make observations. Also calculate the limit to spatial resolution (such as the size of features observable on the Moon) for observations at a specific distance from the device. Among the things to be considered are the wavelength of electromagnetic radiation used, the size of the circular object, and the distance to the system or phenomenon being observed.

Glossary

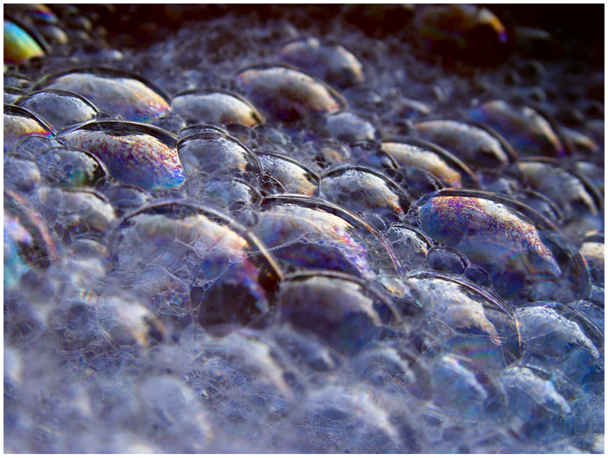
Rayleigh criterion

two images are just resolvable when the center of the diffraction pattern of one is directly over the first minimum of the diffraction pattern of the other

Thin Film Interference

- Discuss the rainbow formation by thin films.

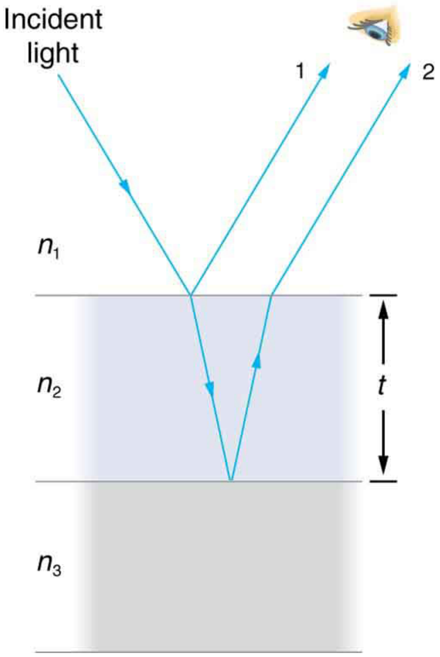
The bright colors seen in an oil slick floating on water or in a sunlit soap bubble are caused by interference. The brightest colors are those that interfere constructively. This interference is between light reflected from different surfaces of a thin film; thus, the effect is known as **thin film interference**. As noticed before, interference effects are most prominent when light interacts with something having a size similar to its wavelength. A thin film is one having a thickness t smaller than a few times the wavelength of light, λ. Since color is associated indirectly with λ and since all interference depends in some way on the ratio of λ to the size of the object involved, we should expect to see different colors for different thicknesses of a film, as in [link].



These soap bubbles exhibit brilliant colors when exposed to sunlight. (credit: Scott Robinson, Flickr)

What causes thin film interference? [link] shows how light reflected from the top and bottom surfaces of a film can interfere. Incident light is only partially reflected from the top surface of the film (ray 1). The remainder enters the film and is itself partially reflected from the bottom surface. Part of the light reflected from the bottom surface can emerge from the top of the film (ray 2) and interfere with light reflected from the top (ray 1). Since the ray that enters the film travels a greater distance, it may be in or out of phase with the ray reflected from the top. However, consider for a moment, again, the bubbles in [link]. The bubbles are darkest where they are thinnest. Furthermore, if you observe a soap bubble carefully, you will note it gets dark at the point where it breaks. For very thin films, the difference in path lengths of ray 1 and ray 2 in [link] is negligible; so why should they interfere destructively and not constructively? The answer is that a phase change can occur upon reflection. The rule is as follows:

When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180° phase change (or a $\lambda/2$ shift) occurs.



Light striking a thin film is partially reflected (ray 1) and partially refracted at the top surface. The refracted ray is partially reflected at the bottom surface and emerges as ray 2. These rays will interfere in a way that depends on the thickness of the film and the indices of refraction of the various media.

If the film in [link] is a soap bubble (essentially water with air on both sides), then there is a $\lambda/2$ shift for ray 1 and none for ray 2. Thus, when the film is very thin, the path length difference between the two rays is negligible, they are exactly out of phase, and destructive interference will occur at all wavelengths and so the soap bubble will be dark here.

The thickness of the film relative to the wavelength of light is the other crucial factor in thin film interference. Ray 2 in [link] travels a greater distance than ray 1. For light incident perpendicular to the surface, ray 2 travels a distance approximately $2t$ farther than ray 1. When this distance is an integral or half-integral multiple of the wavelength in the medium ($\lambda_n = \lambda/n$, where λ is the wavelength in vacuum and n is the index of refraction), constructive or destructive interference occurs, depending also on whether there is a phase change in either ray.

Example:
Calculating Non-reflective Lens Coating Using Thin Film Interference

Sophisticated cameras use a series of several lenses. Light can reflect from the surfaces of these various lenses and degrade image clarity. To limit these reflections, lenses are coated with a thin layer of magnesium fluoride that causes destructive thin film interference. What is the thinnest this film can be, if its index of refraction is 1.38 and it is designed to limit the reflection of 550-nm light, normally the most intense visible wavelength? The index of refraction of glass is 1.52.

Strategy

Refer to [link] and use $n_1 = 100$ for air, $n_2 = 1.38$, and $n_3 = 1.52$. Both ray 1 and ray 2 will have a $\lambda/2$ shift upon reflection. Thus, to obtain destructive interference, ray 2 will need to travel a half wavelength farther than ray 1. For rays incident perpendicularly, the path length difference is $2t$.

Solution

To obtain destructive interference here,

Equation:

$$2t = \frac{\lambda_{n_2}}{2},$$

where λ_{n_2} is the wavelength in the film and is given by $\lambda_{n_2} = \frac{\lambda}{n_2}$.

Thus,

Equation:

$$2t = \frac{\lambda/n_2}{2}.$$

Solving for t and entering known values yields

Equation:

$$\begin{aligned} t &= \frac{\lambda/n_2}{4} = \frac{(550\text{ nm})/1.38}{4} \\ &= 99.6\text{ nm}. \end{aligned}$$

Discussion

Films such as the one in this example are most effective in producing destructive interference when the thinnest layer is used, since light over a broader range of incident angles will be reduced in intensity. These films are called non-reflective coatings; this is only an approximately correct description, though, since other wavelengths will only be partially cancelled. Non-reflective coatings are used in car windows and sunglasses.

Thin film interference is most constructive or most destructive when the path length difference for the two rays is an integral or half-integral wavelength, respectively. That is, for rays incident perpendicularly, $2t = \lambda_n, 2\lambda_n, 3\lambda_n, \ldots$ or $2t = \lambda_n/2, 3\lambda_n/2, 5\lambda_n/2, \ldots$. To know whether interference is constructive or destructive, you must also determine if there is a phase change upon reflection. Thin film interference thus depends on film thickness, the wavelength of light, and the refractive indices. For white light incident on a film that varies in thickness, you will observe rainbow colors of constructive interference for various wavelengths as the thickness varies.

Example:

Soap Bubbles: More Than One Thickness can be Constructive

(a) What are the three smallest thicknesses of a soap bubble that produce constructive interference for red light with a wavelength of 650 nm? The index of refraction of soap is taken to be the same as that of water. (b) What three smallest thicknesses will give destructive interference?

Strategy and Concept

Use [link] to visualize the bubble. Note that $n_1 = n_3 = 1.00$ for air, and $n_2 = 1.333$ for soap (equivalent to water). There is a $\lambda/2$ shift for ray 1 reflected from the top surface of the bubble, and no shift for ray 2 reflected from the bottom surface. To get constructive interference, then, the path length difference ($2t$) must be a half-integral multiple of the wavelength—the first three being $\lambda_n/2$, $3\lambda_n/2$, and $5\lambda_n/2$. To get destructive interference, the path length difference must be an integral multiple of the wavelength—the first three being 0, λ_n, and $2\lambda_n$.

Solution for (a)

Constructive interference occurs here when

Equation:

$$2t_{\mathrm{c}} = \frac{\lambda_n}{2}, \frac{3\lambda_n}{2}, \frac{5\lambda_n}{2}, \ldots.$$

The smallest constructive thickness t_{c} thus is

Equation:

$$\begin{aligned} t_{\text{c}} &= \frac{\lambda_n}{4} = \frac{\lambda/n}{4} = \frac{(650\text{ nm})/1.333}{4} \\ &= 122\text{ nm.} \end{aligned}$$

The next thickness that gives constructive interference is $t\prime_{\text{c}} = 3\lambda_n/4$, so that

Equation:

$$t\prime_{\text{c}} = 366\text{ nm.}$$

Finally, the third thickness producing constructive interference is $t\prime\prime_{\text{c}} \leq 5\lambda_n/4$, so that

Equation:

$$t\prime\prime_{\text{c}} = 610\text{ nm.}$$

Solution for (b)

For *destructive interference*, the path length difference here is an integral multiple of the wavelength. The first occurs for zero thickness, since there is a phase change at the top surface. That is,

Equation:

$$t_{\text{d}} = 0.$$

The first non-zero thickness producing destructive interference is

Equation:

$$2t\prime_{\text{d}} = \lambda_n.$$

Substituting known values gives

Equation:

$$\begin{aligned} t\prime_{\text{d}} &= \frac{\lambda_n}{2} = \frac{\lambda/n}{2} = \frac{(650\text{ nm})/1.333}{2} \\ &= 244\text{ nm.} \end{aligned}$$

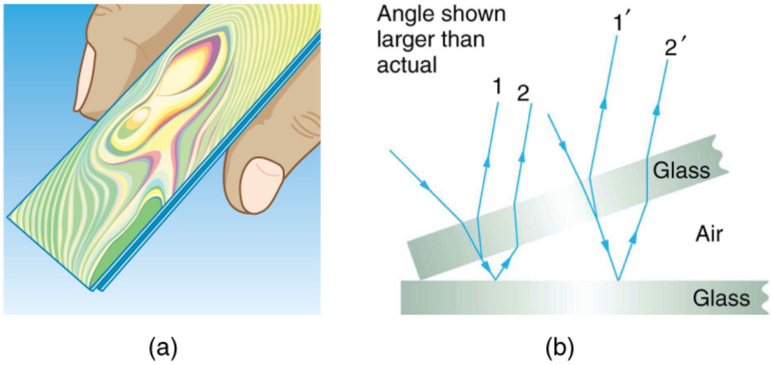
Finally, the third destructive thickness is $2t\prime\prime_{\text{d}} = 2\lambda_n$, so that

Equation:

$$\begin{aligned} t''_{\text{d}} &= \lambda_n = \frac{\lambda}{n} = \frac{650\text{ nm}}{1.333} \\ &= 488\text{ nm.} \end{aligned}$$

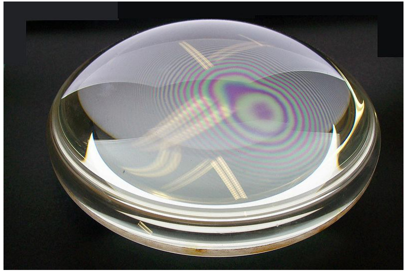
Discussion
If the bubble was illuminated with pure red light, we would see bright and dark bands at very uniform increases in thickness. First would be a dark band at 0 thickness, then bright at 122 nm thickness, then dark at 244 nm, bright at 366 nm, dark at 488 nm, and bright at 610 nm. If the bubble varied smoothly in thickness, like a smooth wedge, then the bands would be evenly spaced.

Another example of thin film interference can be seen when microscope slides are separated (see [link]). The slides are very flat, so that the wedge of air between them increases in thickness very uniformly. A phase change occurs at the second surface but not the first, and so there is a dark band where the slides touch. The rainbow colors of constructive interference repeat, going from violet to red again and again as the distance between the slides increases. As the layer of air increases, the bands become more difficult to see, because slight changes in incident angle have greater effects on path length differences. If pure-wavelength light instead of white light is used, then bright and dark bands are obtained rather than repeating rainbow colors.



(a) The rainbow color bands are produced by thin film interference in the air between the two glass slides. (b) Schematic of the paths taken by rays in the wedge of air between the slides.

An important application of thin film interference is found in the manufacturing of optical instruments. A lens or mirror can be compared with a master as it is being ground, allowing it to be shaped to an accuracy of less than a wavelength over its entire surface. [link] illustrates the phenomenon called Newton’s rings, which occurs when the plane surfaces of two lenses are placed together. (The circular bands are called Newton’s rings because Isaac Newton described them and their use in detail. Newton did not discover them; Robert Hooke did, and Newton did not believe they were due to the wave character of light.) Each successive ring of a given color indicates an increase of only one wavelength in the distance between the lens and the blank, so that great precision can be obtained. Once the lens is perfect, there will be no rings.



“Newton's rings” interference fringes are produced when two plano-convex lenses are placed together with their plane surfaces in contact. The rings are created by interference between the light reflected off the two surfaces as a result of a slight gap between them, indicating that these surfaces are not precisely plane but are slightly convex. (credit: Ulf Seifert, Wikimedia Commons)

The wings of certain moths and butterflies have nearly iridescent colors due to thin film interference. In addition to pigmentation, the wing's color is affected greatly by constructive interference of certain wavelengths reflected from its film-coated surface. Car manufacturers are offering special paint jobs that use thin film interference to produce colors that change with angle. This expensive option is based on variation of thin film path length differences with angle. Security features on credit cards, banknotes, driving licenses and similar items prone to forgery use thin film interference, diffraction gratings, or holograms. Australia led the way with dollar bills printed on polymer with a diffraction grating security feature making the currency difficult to forge. Other countries such as New Zealand and Taiwan are using similar technologies, while the United States currency includes a thin film interference effect.

Note:

Making Connections: Take-Home Experiment—Thin Film Interference

One feature of thin film interference and diffraction gratings is that the pattern shifts as you change the angle at which you look or move your head. Find examples of thin film interference and gratings around you. Explain how the patterns change for each specific example. Find examples where the thickness changes giving rise to changing colors. If you can find two microscope slides, then try observing the effect shown in [link]. Try separating one end of the two slides with a hair or maybe a thin piece of paper and observe the effect.

Problem-Solving Strategies for Wave Optics

Step 1. *Examine the situation to determine that interference is involved*. Identify whether slits or thin film interference are considered in the problem.

Step 2. *If slits are involved*, note that diffraction gratings and double slits produce very similar interference patterns, but that gratings have narrower (sharper) maxima. Single slit patterns are characterized by a large central maximum and smaller maxima to the sides.

Step 3. *If thin film interference is involved, take note of the path length difference between the two rays that interfere*. Be certain to use the wavelength in the medium involved, since it differs from the wavelength in vacuum. Note also that there is an additional $\lambda/2$ phase shift when light reflects from a medium with a greater index of refraction.

Step 4. *Identify exactly what needs to be determined in the problem (identify the unknowns)*. A written list is useful. Draw a diagram of the situation. Labeling the diagram is useful.

Step 5. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns)*.

Step 6. *Solve the appropriate equation for the quantity to be determined (the unknown), and enter the knowns*. Slits, gratings, and the Rayleigh limit involve equations.

Step 7. *For thin film interference, you will have constructive interference for a total shift that is an integral number of wavelengths. You will have destructive interference for a total shift of a half-integral number of wavelengths*. Always keep in mind that crest to crest is constructive whereas crest to trough is destructive.

Step 8. *Check to see if the answer is reasonable: Does it make sense?* Angles in interference patterns cannot be greater than 90º, for example.

Section Summary

- Thin film interference occurs between the light reflected from the top and bottom surfaces of a film. In addition to the path length difference, there can be a phase change.
- When light reflects from a medium having an index of refraction greater than that of the medium in which it is traveling, a 180^{o} phase change (or a $\lambda/2$ shift) occurs.

Conceptual Questions

Exercise:

Problem:

What effect does increasing the wedge angle have on the spacing of interference fringes? If the wedge angle is too large, fringes are not observed. Why?

Exercise:

Problem:

How is the difference in paths taken by two originally in-phase light waves related to whether they interfere constructively or destructively? How can this be affected by reflection? By refraction?

Exercise:

Problem:

Is there a phase change in the light reflected from either surface of a contact lens floating on a person’s tear layer? The index of refraction of the lens is about 1.5, and its top surface is dry.

Exercise:

Problem:

In placing a sample on a microscope slide, a glass cover is placed over a water drop on the glass slide. Light incident from above can reflect from the top and bottom of the glass cover and from the glass slide below the water drop. At which surfaces will there be a phase change in the reflected light?

Exercise:

Problem:

Answer the above question if the fluid between the two pieces of crown glass is carbon disulfide.

Exercise:

Problem:

While contemplating the food value of a slice of ham, you notice a rainbow of color reflected from its moist surface. Explain its origin.

Exercise:

Problem:

An inventor notices that a soap bubble is dark at its thinnest and realizes that destructive interference is taking place for all wavelengths. How could she use this knowledge to make a non-reflective coating for lenses that is effective at all wavelengths? That is, what limits would there be on the index of refraction and thickness of the coating? How might this be impractical?

Exercise:

Problem:

A non-reflective coating like the one described in [link] works ideally for a single wavelength and for perpendicular incidence. What happens for other wavelengths and other incident directions? Be specific.

Exercise:

Problem:

Why is it much more difficult to see interference fringes for light reflected from a thick piece of glass than from a thin film? Would it be easier if monochromatic light were used?

Problems & Exercises

Exercise:

Problem:

A soap bubble is 100 nm thick and illuminated by white light incident perpendicular to its surface. What wavelength and color of visible light is most constructively reflected, assuming the same index of refraction as water?

Solution:

532 nm (green)

Exercise:

Problem:

An oil slick on water is 120 nm thick and illuminated by white light incident perpendicular to its surface. What color does the oil appear (what is the most constructively reflected wavelength), given its index of refraction is 1.40?

Exercise:

Problem:

Calculate the minimum thickness of an oil slick on water that appears blue when illuminated by white light perpendicular to its surface. Take the blue wavelength to be 470 nm and the index of refraction of oil to be 1.40.

Solution:

83.9 nm

Exercise:

Problem:

Find the minimum thickness of a soap bubble that appears red when illuminated by white light perpendicular to its surface. Take the wavelength to be 680 nm, and assume the same index of refraction as water.

Exercise:

Problem:

A film of soapy water ($n = 1.33$) on top of a plastic cutting board has a thickness of 233 nm. What color is most strongly reflected if it is illuminated perpendicular to its surface?

Solution:

620 nm (orange)

Exercise:

Problem:

What are the three smallest non-zero thicknesses of soapy water ($n = 1.33$) on Plexiglas if it appears green (constructively reflecting 520-nm light) when illuminated perpendicularly by white light? Explicitly show how you follow the steps in Problem Solving Strategies for Wave Optics.

Exercise:

Problem:

Suppose you have a lens system that is to be used primarily for 700-nm red light. What is the second thinnest coating of fluorite (magnesium fluoride) that would be non-reflective for this wavelength?

Solution:

380 nm

Exercise:

Problem:

(a) As a soap bubble thins it becomes dark, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the bubble can be and appear dark at all visible wavelengths? Assume the same index of refraction as water. (b) Discuss the fragility of the film considering the thickness found.

Exercise:

Problem:

A film of oil on water will appear dark when it is very thin, because the path length difference becomes small compared with the wavelength of light and there is a phase shift at the top surface. If it becomes dark when the path length difference is less than one-fourth the wavelength, what is the thickest the oil can be and appear dark at all visible wavelengths? Oil has an index of refraction of 1.40.

Solution:

33.9 nm

Exercise:

Problem:

[link] shows two glass slides illuminated by pure-wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on a 0.100-mm-diameter hair at the other end, forming a wedge of air. (a) How far apart are the dark bands, if the slides are 7.50 cm long and 589-nm light is used? (b) Is there any difference if the slides are made from crown or flint glass? Explain.

Exercise:

Problem:

[link] shows two 7.50-cm-long glass slides illuminated by pure 589-nm wavelength light incident perpendicularly. The top slide touches the bottom slide at one end and rests on some debris at the other end, forming a wedge of air. How thick is the debris, if the dark bands are 1.00 mm apart?

Solution:

$4.42\times 10^{-5}\text{ m}$

Exercise:

Problem: Repeat [link], but take the light to be incident at a 45^{o} angle.

Exercise:

Problem: Repeat [link], but take the light to be incident at a 45^{o} angle.

Solution:

The oil film will appear black, since the reflected light is not in the visible part of the spectrum.

Exercise:

Problem: **Unreasonable Results**

To save money on making military aircraft invisible to radar, an inventor decides to coat them with a non-reflective material having an index of refraction of 1.20, which is between that of air and the surface of the plane. This, he reasons, should be much cheaper than designing Stealth bombers. (a) What thickness should the coating be to inhibit the reflection of 4.00-cm wavelength radar? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

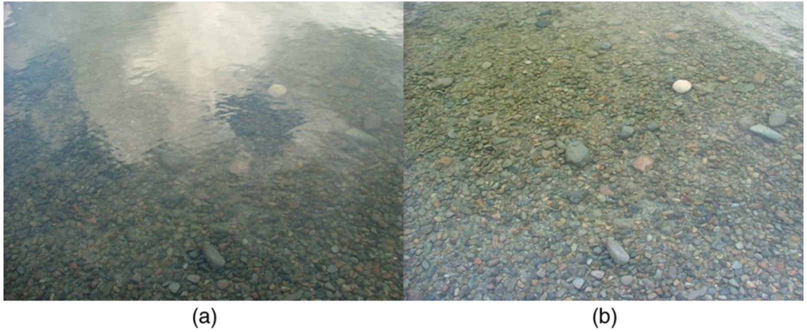
thin film interference

interference between light reflected from different surfaces of a thin film

Polarization

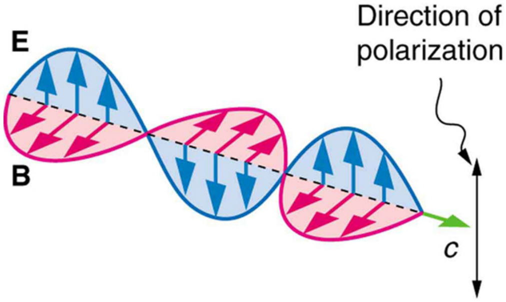
- Discuss the meaning of polarization.
- Discuss the property of optical activity of certain materials.

Polaroid sunglasses are familiar to most of us. They have a special ability to cut the glare of light reflected from water or glass (see [link]). Polaroids have this ability because of a wave characteristic of light called polarization. What is polarization? How is it produced? What are some of its uses? The answers to these questions are related to the wave character of light.



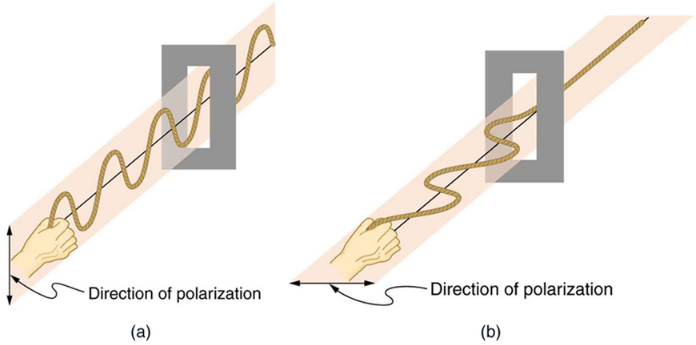
These two photographs of a river show the effect of a polarizing filter in reducing glare in light reflected from the surface of water. Part (b) of this figure was taken with a polarizing filter and part (a) was not. As a result, the reflection of clouds and sky observed in part (a) is not observed in part (b). Polarizing sunglasses are particularly useful on snow and water. (credit: Amithshs, Wikimedia Commons)

Light is one type of electromagnetic (EM) wave. As noted earlier, EM waves are *transverse waves* consisting of varying electric and magnetic fields that oscillate perpendicular to the direction of propagation (see [link]). There are specific directions for the oscillations of the electric and magnetic fields. **Polarization** is the attribute that a wave’s oscillations have a definite direction relative to the direction of propagation of the wave. (This is not the same type of polarization as that discussed for the separation of charges.) Waves having such a direction are said to be **polarized**. For an EM wave, we define the **direction of polarization** to be the direction parallel to the electric field. Thus we can think of the electric field arrows as showing the direction of polarization, as in [link].



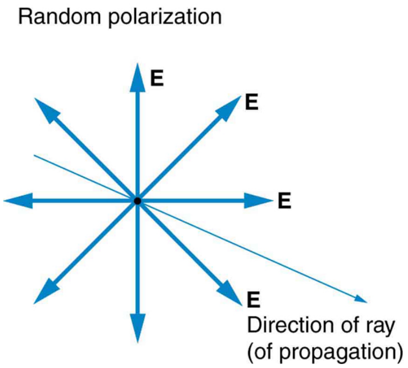
An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

To examine this further, consider the transverse waves in the ropes shown in [link]. The oscillations in one rope are in a vertical plane and are said to be **vertically polarized**. Those in the other rope are in a horizontal plane and are **horizontally polarized**. If a vertical slit is placed on the first rope, the waves pass through. However, a vertical slit blocks the horizontally polarized waves. For EM waves, the direction of the electric field is analogous to the disturbances on the ropes.

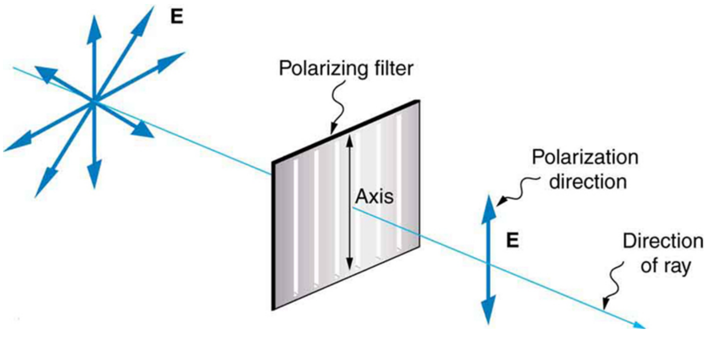


The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally polarized waves.

The Sun and many other light sources produce waves that are randomly polarized (see [link]). Such light is said to be **unpolarized** because it is composed of many waves with all possible directions of polarization. Polaroid materials, invented by the founder of Polaroid Corporation, Edwin Land, act as a *polarizing* slit for light, allowing only polarization in one direction to pass through. Polarizing filters are composed of long molecules aligned in one direction. Thinking of the molecules as many slits, analogous to those for the oscillating ropes, we can understand why only light with a specific polarization can get through. The **axis of a polarizing filter** is the direction along which the filter passes the electric field of an EM wave (see [link]).



The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.



A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM

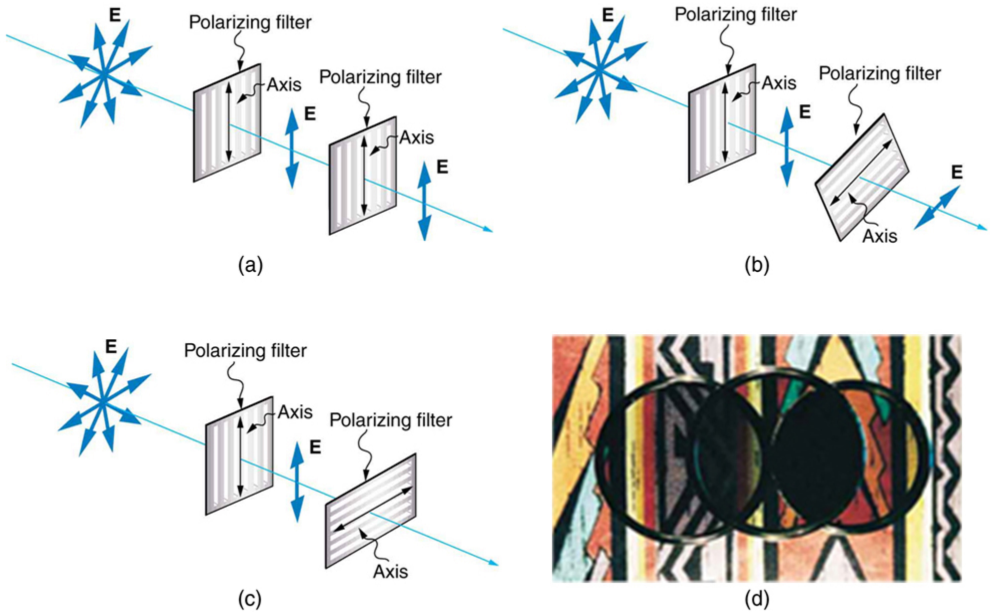
wave is defined to be the direction of
its electric field.

[link] shows the effect of two polarizing filters on originally unpolarized light. The first filter polarizes the light along its axis. When the axes of the first and second filters are aligned (parallel), then all of the polarized light passed by the first filter is also passed by the second. If the second polarizing filter is rotated, only the component of the light parallel to the second filter’s axis is passed. When the axes are perpendicular, no light is passed by the second.

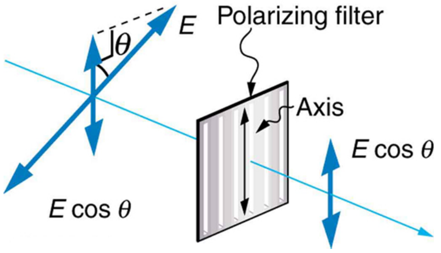
Only the component of the EM wave parallel to the axis of a filter is passed. Let us call the angle between the direction of polarization and the axis of a filter θ. If the electric field has an amplitude E, then the transmitted part of the wave has an amplitude $E\cos\theta$ (see [link]). Since the intensity of a wave is proportional to its amplitude squared, the intensity I of the transmitted wave is related to the incident wave by
Equation:

$$I = I_0 \cos^2 \theta,$$

where I_0 is the intensity of the polarized wave before passing through the filter. (The above equation is known as Malus’s law.)



The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is perpendicular to the first, no light is passed. (d) In this photograph, a polarizing filter is placed above two others. Its axis is perpendicular to the filter on the right (dark area) and parallel to the filter on the left (lighter area). (credit: P.P. Urone)



A polarizing filter transmits only the component of the wave parallel to its axis, $E \cos \theta$, reducing the intensity of any light not polarized parallel to its axis.

Example:
Calculating Intensity Reduction by a Polarizing Filter

What angle is needed between the direction of polarized light and the axis of a polarizing filter to reduce its intensity by 90.0%?

Strategy

When the intensity is reduced by 90.0%, it is 10.0% or 0.100 times its original value. That is, $I = 0.100 I_0$. Using this information, the equation $I = I_0 \cos^2 \theta$ can be used to solve for the needed angle.

Solution

Solving the equation $I = I_0 \cos^2 \theta$ for $\cos \theta$ and substituting with the relationship between I and I_0 gives

Equation:

$$\cos \theta = \sqrt{\frac{I}{I_0}} = \sqrt{\frac{0.100 I_0}{I_0}} = 0.3162.$$

Solving for θ yields

Equation:

$$\theta = \cos^{-1} 0.3162 = 71.6^{\mathrm{o}}.$$

Discussion

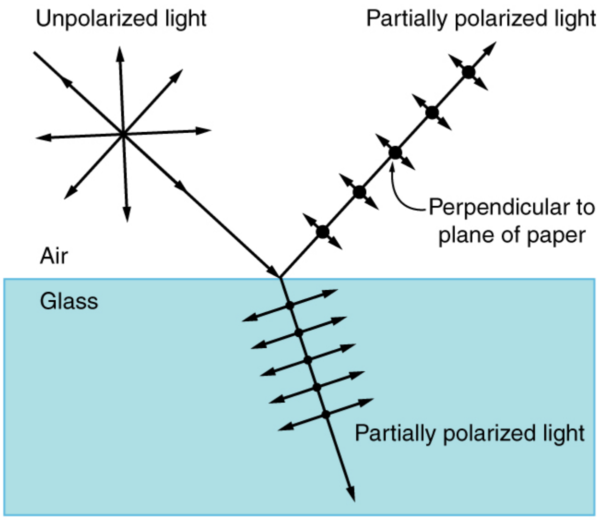
A fairly large angle between the direction of polarization and the filter axis is needed to reduce the intensity to 10.0% of its original value. This seems reasonable based on experimenting with polarizing films. It is interesting that, at an angle of 45^{o}, the intensity is reduced to 50% of its original value (as you will show in this section’s Problems & Exercises). Note that 71.6^{o}

is 18.4º from reducing the intensity to zero, and that at an angle of 18.4º the intensity is reduced to 90.0% of its original value (as you will also show in Problems & Exercises), giving evidence of symmetry.

Polarization by Reflection

By now you can probably guess that Polaroid sunglasses cut the glare in reflected light because that light is polarized. You can check this for yourself by holding Polaroid sunglasses in front of you and rotating them while looking at light reflected from water or glass. As you rotate the sunglasses, you will notice the light gets bright and dim, but not completely black. This implies the reflected light is partially polarized and cannot be completely blocked by a polarizing filter.

[link] illustrates what happens when unpolarized light is reflected from a surface. Vertically polarized light is preferentially refracted at the surface, so that *the reflected light is left more horizontally polarized*. The reasons for this phenomenon are beyond the scope of this text, but a convenient mnemonic for remembering this is to imagine the polarization direction to be like an arrow. Vertical polarization would be like an arrow perpendicular to the surface and would be more likely to stick and not be reflected. Horizontal polarization is like an arrow bouncing on its side and would be more likely to be reflected. Sunglasses with vertical axes would then block more reflected light than unpolarized light from other sources.


Unpolarized light
Partially polarized light
Perpendicular to plane of paper
Air
Glass
Partially polarized light

Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking on their sides bouncing off, whereas arrows striking on their tips go into the surface.

Since the part of the light that is not reflected is refracted, the amount of polarization depends on the indices of refraction of the media involved. It can be shown that **reflected light is completely polarized** at a angle of reflection θ_b, given by

Equation:

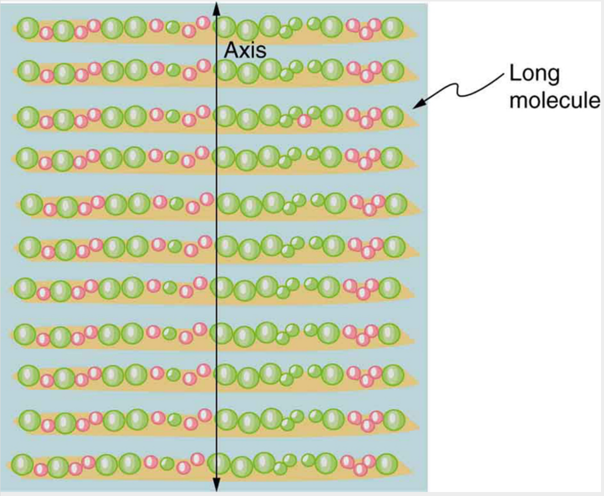
$$\tan\theta_\mathrm{b} = \frac{n_2}{n_1},$$

where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light. This equation is known as **Brewster’s law**, and θ_b is known as **Brewster’s angle**, named after the 19th-century Scottish physicist who discovered them.

Note:

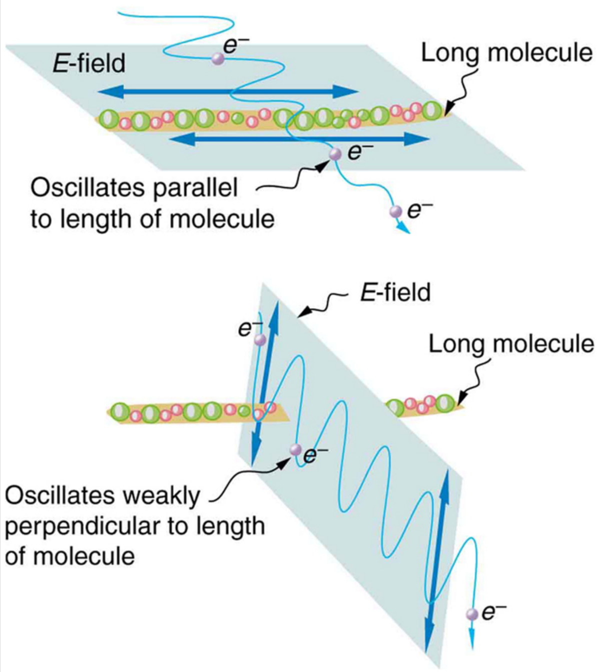
Things Great and Small: Atomic Explanation of Polarizing Filters

Polarizing filters have a polarization axis that acts as a slit. This slit passes electromagnetic waves (often visible light) that have an electric field parallel to the axis. This is accomplished with long molecules aligned perpendicular to the axis as shown in [link].



Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

[link] illustrates how the component of the electric field parallel to the long molecules is absorbed. An electromagnetic wave is composed of oscillating electric and magnetic fields. The electric field is strong compared with the magnetic field and is more effective in exerting force on charges in the molecules. The most affected charged particles are the electrons in the molecules, since electron masses are small. If the electron is forced to oscillate, it can absorb energy from the EM wave. This reduces the fields in the wave and, hence, reduces its intensity. In long molecules, electrons can more easily oscillate parallel to the molecule than in the perpendicular direction. The electrons are bound to the molecule and are more restricted in their movement perpendicular to the molecule. Thus, the electrons can absorb EM waves that have a component of their electric field parallel to the molecule. The electrons are much less responsive to electric fields perpendicular to the molecule and will allow those fields to pass. Thus the axis of the polarizing filter is perpendicular to the length of the molecule.



e^-
E-field
Long molecule
e^-
Oscillates parallel to length of molecule
e^-
E-field
e^-
Long molecule
e^-
Oscillates weakly perpendicular to length of molecule
e^-

Artist’s conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

Example:
Calculating Polarization by Reflection
(a) At what angle will light traveling in air be completely polarized horizontally when reflected from water? (b) From glass?

Strategy

All we need to solve these problems are the indices of refraction. Air has $n_1 = 1.00$, water has $n_2 = 1.333$, and crown glass has $n\prime_2 = 1.520$. The equation $\tan\,\theta_\text{b} = \frac{n_2}{n_1}$ can be directly applied to find θ_b in each case.

Solution for (a)

Putting the known quantities into the equation

Equation:

$$\tan\,\theta_\text{b} = \frac{n_2}{n_1}$$

gives

Equation:

$$\tan\,\theta_\text{b} = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333.$$

Solving for the angle θ_b yields

Equation:

$$\theta_\text{b} = \tan^{-1}1.333 = 53.1^\text{o}.$$

Solution for (b)

Similarly, for crown glass and air,

Equation:

$$\tan \theta\prime_{\text{b}} = \frac{n\prime_2}{n_1} = \frac{1.520}{1.00} = 1.52.$$

Thus,

Equation:

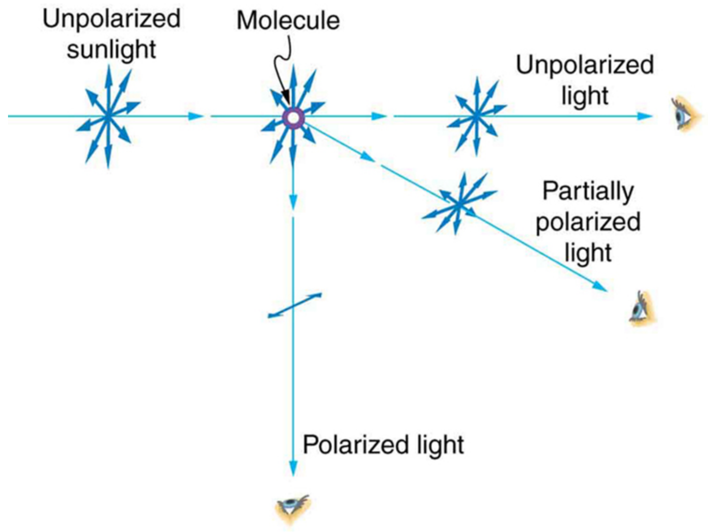
$$\theta\prime_{\text{b}} = \tan^{-1} 1.52 = 56.7^{\text{o}}.$$

Discussion

Light reflected at these angles could be completely blocked by a good polarizing filter held with its *axis vertical*. Brewster's angle for water and air are similar to those for glass and air, so that sunglasses are equally effective for light reflected from either water or glass under similar circumstances. Light not reflected is refracted into these media. So at an incident angle equal to Brewster's angle, the refracted light will be slightly polarized vertically. It will not be completely polarized vertically, because only a small fraction of the incident light is reflected, and so a significant amount of horizontally polarized light is refracted.

Polarization by Scattering

If you hold your Polaroid sunglasses in front of you and rotate them while looking at blue sky, you will see the sky get bright and dim. This is a clear indication that light scattered by air is partially polarized. [link] helps illustrate how this happens. Since light is a transverse EM wave, it vibrates the electrons of air molecules perpendicular to the direction it is traveling. The electrons then radiate like small antennae. Since they are oscillating perpendicular to the direction of the light ray, they produce EM radiation that is polarized perpendicular to the direction of the ray. When viewing the light along a line perpendicular to the original ray, as in [link], there can be no polarization in the scattered light parallel to the original ray, because that would require the original ray to be a longitudinal wave. Along other directions, a component of the other polarization can be projected along the line of sight, and the scattered light will only be partially polarized. Furthermore, multiple scattering can bring light to your eyes from other directions and can contain different polarizations.



Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

Photographs of the sky can be darkened by polarizing filters, a trick used by many photographers to make clouds brighter by contrast. Scattering from other particles, such as smoke or dust, can also polarize light. Detecting polarization in scattered EM waves can be a useful analytical tool in determining the scattering source.

There is a range of optical effects used in sunglasses. Besides being Polaroid, other sunglasses have colored pigments embedded in them, while others use non-reflective or even reflective coatings. A recent development is photochromic lenses, which darken in the sunlight and become clear indoors. Photochromic lenses are embedded with organic microcrystalline molecules that change their properties when exposed to UV in sunlight, but become clear in artificial lighting with no UV.

Note:

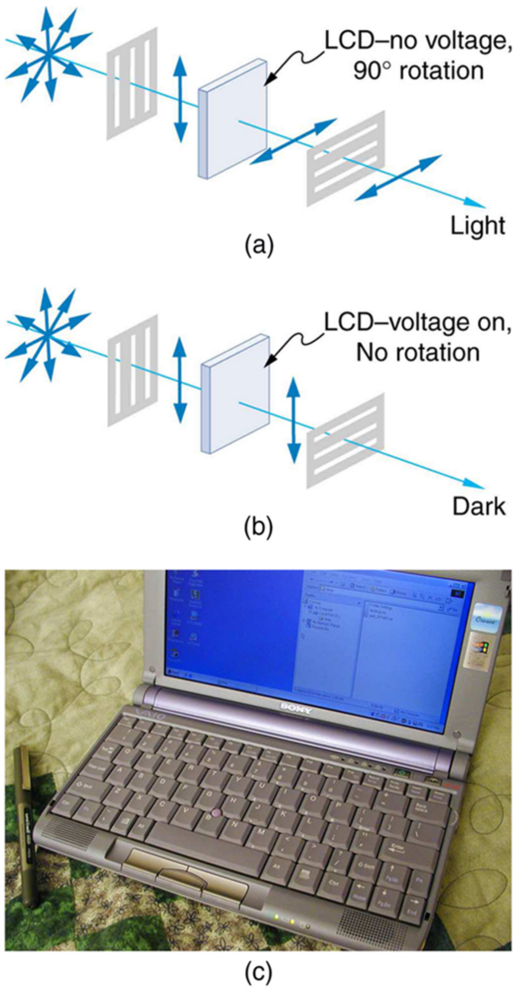
Take-Home Experiment: Polarization

Find Polaroid sunglasses and rotate one while holding the other still and look at different surfaces and objects. Explain your observations. What is the difference in angle from when you see a maximum intensity to when you see a minimum intensity? Find a reflective glass surface and do the same. At what angle does the glass need to be oriented to give minimum glare?

Liquid Crystals and Other Polarization Effects in Materials

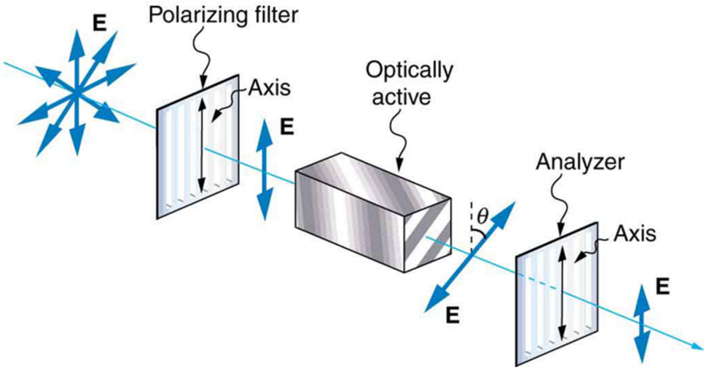
While you are undoubtedly aware of liquid crystal displays (LCDs) found in watches, calculators, computer screens, cellphones, flat screen televisions, and other myriad places, you may not be aware that they are based on polarization. Liquid crystals are so named because their molecules can be aligned even though they are in a liquid. Liquid crystals have the property that they can rotate the polarization of light passing through them by 90^o. Furthermore, this property can be turned off by the application of a voltage, as illustrated in [link]. It is possible to manipulate this characteristic quickly and in small well-defined regions to create the contrast patterns we see in so many LCD devices.

In flat screen LCD televisions, there is a large light at the back of the TV. The light travels to the front screen through millions of tiny units called pixels (picture elements). One of these is shown in [link] (a) and (b). Each unit has three cells, with red, blue, or green filters, each controlled independently. When the voltage across a liquid crystal is switched off, the liquid crystal passes the light through the particular filter. One can vary the picture contrast by varying the strength of the voltage applied to the liquid crystal.



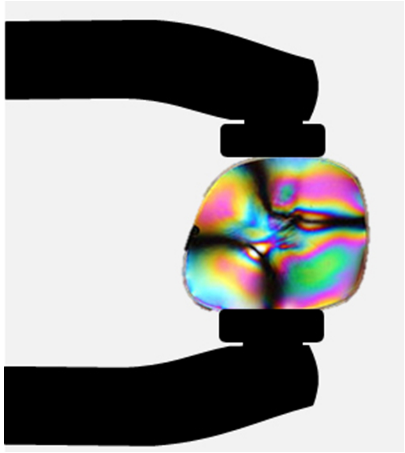
(a) Polarized light is rotated 90º by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making the region dark in comparison with its surroundings. (c) LCDs can be made color specific, small, and fast enough to use in laptop computers and TVs. (credit: Jon Sullivan)

Many crystals and solutions rotate the plane of polarization of light passing through them. Such substances are said to be **optically active**. Examples include sugar water, insulin, and collagen (see [link]). In addition to depending on the type of substance, the amount and direction of rotation depends on a number of factors. Among these is the concentration of the substance, the distance the light travels through it, and the wavelength of light. Optical activity is due to the asymmetric shape of molecules in the substance, such as being helical. Measurements of the rotation of polarized light passing through substances can thus be used to measure concentrations, a standard technique for sugars. It can also give information on the shapes of molecules, such as proteins, and factors that affect their shapes, such as temperature and pH.



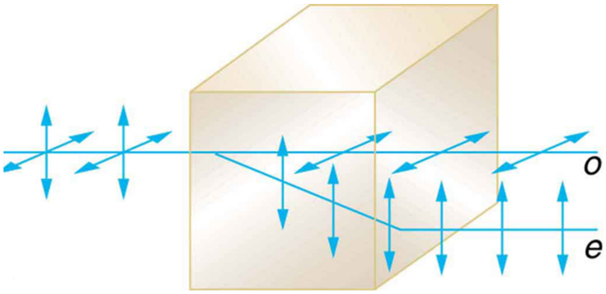
Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

Glass and plastic become optically active when stressed; the greater the stress, the greater the effect. Optical stress analysis on complicated shapes can be performed by making plastic models of them and observing them through crossed filters, as seen in [link]. It is apparent that the effect depends on wavelength as well as stress. The wavelength dependence is sometimes also used for artistic purposes.



Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

Another interesting phenomenon associated with polarized light is the ability of some crystals to split an unpolarized beam of light into two. Such crystals are said to be **birefringent** (see [link]). Each of the separated rays has a specific polarization. One behaves normally and is called the ordinary ray, whereas the other does not obey Snell’s law and is called the extraordinary ray. Birefringent crystals can be used to produce polarized beams from unpolarized light. Some birefringent materials preferentially absorb one of the polarizations. These materials are called dichroic and can produce polarization by this preferential absorption. This is fundamentally how polarizing filters and other polarizers work. The interested reader is invited to further pursue the numerous properties of materials related to polarization.



Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell’s law.

Section Summary

- Polarization is the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave.
- EM waves are transverse waves that may be polarized.
- The direction of polarization is defined to be the direction parallel to the electric field of the EM wave.
- Unpolarized light is composed of many rays having random polarization directions.

- Light can be polarized by passing it through a polarizing filter or other polarizing material. The intensity I of polarized light after passing through a polarizing filter is $I = I_0 \cos^2$ θ, where I_0 is the original intensity and θ is the angle between the direction of polarization and the axis of the filter.
- Polarization is also produced by reflection.
- Brewster’s law states that reflected light will be completely polarized at the angle of reflection θ_{b}, known as Brewster’s angle, given by a statement known as Brewster’s law: $\tan \theta_{\text{b}} = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light.
- Polarization can also be produced by scattering.
- There are a number of types of optically active substances that rotate the direction of polarization of light passing through them.

Conceptual Questions

Exercise:

Problem:

Under what circumstances is the phase of light changed by reflection? Is the phase related to polarization?

Exercise:

Problem: Can a sound wave in air be polarized? Explain.

Exercise:

Problem:

No light passes through two perfect polarizing filters with perpendicular axes. However, if a third polarizing filter is placed between the original two, some light can pass. Why is this? Under what circumstances does most of the light pass?

Exercise:

Problem:

Explain what happens to the energy carried by light that it is dimmed by passing it through two crossed polarizing filters.

Exercise:

Problem:

When particles scattering light are much smaller than its wavelength, the amount of scattering is proportional to $1/\lambda^4$. Does this mean there is more scattering for small λ than large λ? How does this relate to the fact that the sky is blue?

Exercise:

Problem:

Using the information given in the preceding question, explain why sunsets are red.

Exercise:

Problem:

When light is reflected at Brewster’s angle from a smooth surface, it is 100% polarized parallel to the surface. Part of the light will be refracted into the surface. Describe how you would do an experiment to determine the polarization of the refracted light. What direction would you expect the polarization to have and would you expect it to be 100%?

Problems & Exercises

Exercise:

Problem:

What angle is needed between the direction of polarized light and the axis of a polarizing filter to cut its intensity in half?

Solution:

45.0°

Exercise:

Problem:

The angle between the axes of two polarizing filters is 45.0°. By how much does the second filter reduce the intensity of the light coming through the first?

Exercise:

Problem:

If you have completely polarized light of intensity $150\ \mathrm{W/m^2}$, what will its intensity be after passing through a polarizing filter with its axis at an 89.0° angle to the light's polarization direction?

Solution:

$45.7\ \mathrm{mW/m^2}$

Exercise:

Problem:

What angle would the axis of a polarizing filter need to make with the direction of polarized light of intensity $1.00\ \mathrm{kW/m^2}$ to reduce the intensity to $10.0\ \mathrm{W/m^2}$?

Exercise:

Problem:

At the end of [link], it was stated that the intensity of polarized light is reduced to 90.0% of its original value by passing through a polarizing filter with its axis at an angle of 18.4° to the direction of polarization. Verify this statement.

Solution:

90.0%

Exercise:

Problem:

Show that if you have three polarizing filters, with the second at an angle of 45^{o} to the first and the third at an angle of 90.0^{o} to the first, the intensity of light passed by the first will be reduced to 25.0% of its value. (This is in contrast to having only the first and third, which reduces the intensity to zero, so that placing the second between them increases the intensity of the transmitted light.)

Exercise:

Problem:

Prove that, if I is the intensity of light transmitted by two polarizing filters with axes at an angle θ and $I\prime$ is the intensity when the axes are at an angle $90.0^{\text{o}} - \text{θ}$, then $I + I\prime = I_0$, the original intensity. (Hint: Use the trigonometric identities $\cos\,(90.0^{\text{o}} - \theta) = \sin\,\theta$ and $\cos^2\,\theta + \sin^2\,\theta = 1$.)

Solution:

I_0

Exercise:

Problem:

At what angle will light reflected from diamond be completely polarized?

Exercise:

Problem:

What is Brewster’s angle for light traveling in water that is reflected from crown glass?

Solution:

48.8^o

Exercise:

Problem:

A scuba diver sees light reflected from the water’s surface. At what angle will this light be completely polarized?

Exercise:

Problem:

At what angle is light inside crown glass completely polarized when reflected from water, as in a fish tank?

Solution:

41.2^o

Exercise:

Problem:

Light reflected at 55.6^o from a window is completely polarized. What is the window’s index of refraction and the likely substance of which it is made?

Exercise:

Problem:

(a) Light reflected at 62.5^o from a gemstone in a ring is completely polarized. Can the gem be a diamond? (b) At what angle would the light be completely polarized if the gem was in water?

Solution:

(a) 1.92, not diamond (Zircon)

(b) 55.2^o

Exercise:

Problem:

If θ_{b} is Brewster’s angle for light reflected from the top of an interface between two substances, and $\theta\prime_{\text{b}}$ is Brewster’s angle for light reflected from below, prove that $\theta_{\text{b}} + \theta\prime_{\text{b}} = 90.0^{\text{o}}$.

Exercise:

Problem: Integrated Concepts

If a polarizing filter reduces the intensity of polarized light to 50.0% of its original value, by how much are the electric and magnetic fields reduced?

Solution:

$B_2 = 0.707\ B_1$

Exercise:

Problem: Integrated Concepts

Suppose you put on two pairs of Polaroid sunglasses with their axes at an angle of 15.0^{o}. How much longer will it take the light to deposit a given amount of energy in your eye compared with a single pair of sunglasses? Assume the lenses are clear except for their polarizing characteristics.

Exercise:

Problem: Integrated Concepts

(a) On a day when the intensity of sunlight is $1.00\ \text{kW/m}^2$, a circular lens 0.200 m in diameter focuses light onto water in a black beaker. Two polarizing sheets of plastic are placed in front of the lens with their axes at an angle of 20.0^{o}. Assuming the sunlight is unpolarized and the polarizers are 100% efficient, what is the initial rate of heating of the water in $^{\text{o}}\text{C/s}$, assuming it is 80.0% absorbed? The aluminum beaker has a mass of 30.0 grams and contains 250 grams of water. (b) Do the polarizing filters get hot? Explain.

Solution:

(a) 2.07×10^{-2} °C/s

(b) Yes, the polarizing filters get hot because they absorb some of the lost energy from the sunlight.

Glossary

axis of a polarizing filter
: the direction along which the filter passes the electric field of an EM wave

birefringent
: crystals that split an unpolarized beam of light into two beams

Brewster’s angle
: $\theta_{\mathrm{b}} = \tan^{-1} \frac{n_2}{n_1}$ where n_2 is the index of refraction of the medium from which the light is reflected and n_1 is the index of refraction of the medium in which the reflected light travels

Brewster’s law
: $\tan \theta_{\mathrm{b}} = \frac{n_2}{n_1}$, where n_1 is the medium in which the incident and reflected light travel and n_2 is the index of refraction of the medium that forms the interface that reflects the light

direction of polarization
: the direction parallel to the electric field for EM waves

horizontally polarized
: the oscillations are in a horizontal plane

optically active

substances that rotate the plane of polarization of light passing through them

polarization
the attribute that wave oscillations have a definite direction relative to the direction of propagation of the wave

polarized
waves having the electric and magnetic field oscillations in a definite direction

reflected light that is completely polarized
light reflected at the angle of reflection θ_b, known as Brewster’s angle

unpolarized
waves that are randomly polarized

vertically polarized
the oscillations are in a vertical plane

Extended Topic Microscopy Enhanced by the Wave Characteristics of Light

- Discuss the different types of microscopes.

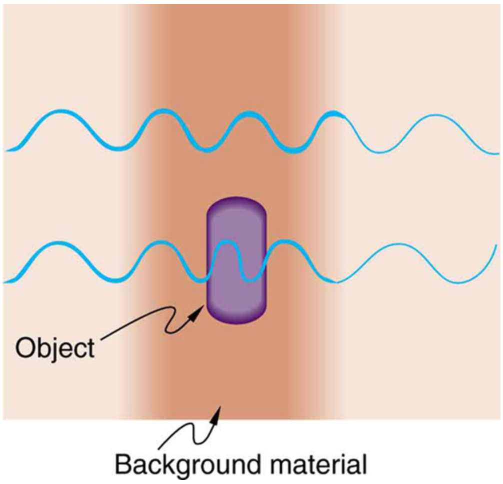
Physics research underpins the advancement of developments in microscopy. As we gain knowledge of the wave nature of electromagnetic waves and methods to analyze and interpret signals, new microscopes that enable us to “see” more are being developed. It is the evolution and newer generation of microscopes that are described in this section.

The use of microscopes (microscopy) to observe small details is limited by the wave nature of light. Owing to the fact that light diffracts significantly around small objects, it becomes impossible to observe details significantly smaller than the wavelength of light. One rule of thumb has it that all details smaller than about are difficult to observe. Radar, for example, can detect the size of an aircraft, but not its individual rivets, since the wavelength of most radar is several centimeters or greater. Similarly, visible light cannot detect individual atoms, since atoms are about 0.1 nm in size and visible wavelengths range from 380 to 760 nm. Ironically, special techniques used to obtain the best possible resolution with microscopes take advantage of the same wave characteristics of light that ultimately limit the detail.

Note:
Making Connections: Waves
All attempts to observe the size and shape of objects are limited by the wavelength of the probe. Sonar and medical ultrasound are limited by the wavelength of sound they employ. We shall see that this is also true in electron microscopy, since electrons have a wavelength. Heisenberg’s uncertainty principle asserts that this limit is fundamental and inescapable, as we shall see in quantum mechanics.

The most obvious method of obtaining better detail is to utilize shorter wavelengths. **Ultraviolet (UV) microscopes** have been constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images. The shorter UV wavelengths allow somewhat greater detail to be observed, but drawbacks, such as the hazard of UV to living tissue and the need for special detection devices and lenses (which tend to be dispersive in the UV), severely limit the use of UV microscopes. Elsewhere, we will explore practical uses of very short wavelength EM waves, such as x rays, and other short-wavelength probes, such as electrons in electron microscopes, to detect small details.

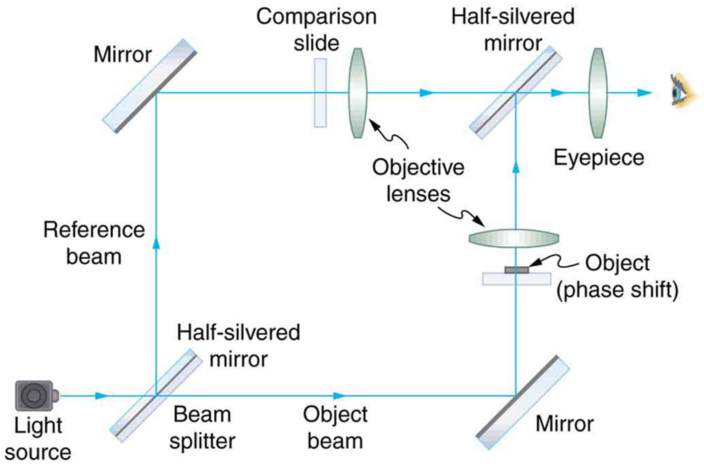
Another difficulty in microscopy is the fact that many microscopic objects do not absorb much of the light passing through them. The lack of contrast makes image interpretation very difficult. **Contrast** is the difference in intensity between objects and the background on which they are observed. Stains (such as dyes, fluorophores, etc.) are commonly employed to enhance contrast, but these tend to be application specific. More general wave interference techniques can be used to produce contrast. [link] shows the passage of light through a sample. Since the indices of refraction differ, the number of wavelengths in the paths differs. Light emerging from the object is thus out of phase with light from the background and will interfere differently, producing enhanced contrast, especially if the light is coherent and monochromatic—as in laser light.



Light rays passing through a sample under a microscope will emerge with different phases depending on their paths.

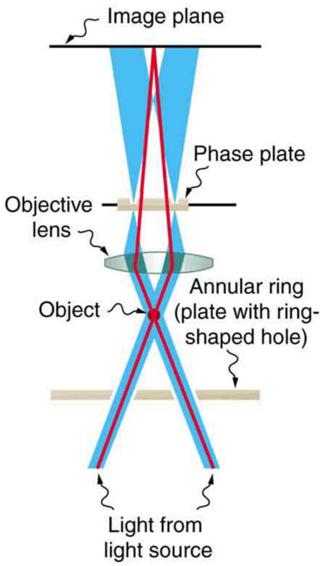
The object shown has a greater index of refraction than the background, and so the wavelength decreases as the ray passes through it. Superimposing these rays produces interference that varies with path, enhancing contrast between the object and background.

Interference microscopes enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample. Since light from the background and objects differ in phase, there will be different amounts of constructive and destructive interference, producing the desired contrast in final intensity. [link] shows schematically how this is done. Parallel rays of light from a source are split into two beams by a half-silvered mirror. These beams are called the object and reference beams. Each beam passes through identical optical elements, except that the object beam passes through the object we wish to observe microscopically. The light beams are recombined by another half-silvered mirror and interfere. Since the light rays passing through different parts of the object have different phases, interference will be significantly different and, hence, have greater contrast between them.



An interference microscope utilizes interference between the reference and object beam to enhance contrast. The two beams are split by a half-silvered mirror; the object beam is sent through the object, and the reference beam is sent through otherwise identical optical elements. The beams are recombined by another half-silvered mirror, and the interference depends on the various phases emerging from different parts of the object, enhancing contrast.

Another type of microscope utilizing wave interference and differences in phases to enhance contrast is called the **phase-contrast microscope**. While its principle is the same as the interference microscope, the phase-contrast microscope is simpler to use and construct. Its impact (and the principle upon which it is based) was so important that its developer, the Dutch physicist Frits Zernike (1888–1966), was awarded the Nobel Prize in 1953. [link] shows the basic construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate (so called because it shifts the phase of the light passing through it). These two light rays are superimposed in the image plane, producing contrast due to their interference.

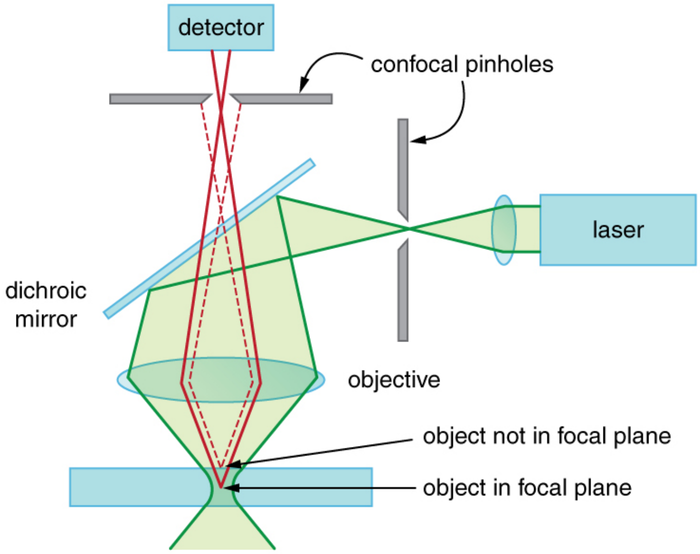


Simplified construction of a phase-contrast microscope. Phase differences between light passing through the object and background are produced by passing the rays through different parts of a phase plate. The light rays are superimposed in the image plane, producing contrast due to their interference.

A **polarization microscope** also enhances contrast by utilizing a wave characteristic of light. Polarization microscopes are useful for objects that are optically active or birefringent, particularly if those characteristics vary from place to place in the object. Polarized light is sent through the object and then observed through a polarizing filter that is perpendicular to the original polarization direction. Nearly transparent objects can then appear with strong color and in high contrast. Many polarization effects are wavelength dependent, producing color in the processed image. Contrast results from the action of the polarizing filter in passing only components parallel to its axis.

Apart from the UV microscope, the variations of microscopy discussed so far in this section are available as attachments to fairly standard microscopes or as slight variations. The next level of sophistication is provided by commercial **confocal microscopes**, which use the extended focal region shown in [link](b) to obtain three-dimensional images rather than two-dimensional images. Here, only a single plane or region of focus is identified; out-of-focus regions above and below this plane are subtracted out by a computer so the image quality is much better. This type of microscope makes use of fluorescence, where a laser provides the excitation light. Laser light passing through a tiny aperture called a pinhole forms an extended focal region within the specimen. The reflected light passes through the objective lens to a second pinhole and the photomultiplier detector, see [link]. The second pinhole is the key here and serves to block much of the light from points that are not at the focal point of the objective lens. The pinhole is conjugate (coupled) to the focal point of the lens. The second pinhole and detector are scanned, allowing reflected light from a small region or section of the extended focal region to be imaged at any one time. The out-of-focus light is excluded. Each image is stored in a computer, and a full scanned image is generated in a short time. Live cell processes can also be imaged at adequate scanning speeds allowing the imaging of three-dimensional microscopic movement. Confocal microscopy enhances images over conventional optical microscopy, especially for thicker specimens, and so has become quite popular.

The next level of sophistication is provided by microscopes attached to instruments that isolate and detect only a small wavelength band of light—monochromators and spectral analyzers. Here, the monochromatic light from a laser is scattered from the specimen. This scattered light shifts up or down as it excites particular energy levels in the sample. The uniqueness of the observed scattered light can give detailed information about the chemical composition of a given spot on the sample with high contrast—like molecular fingerprints. Applications are in materials science, nanotechnology, and the biomedical field. Fine details in biochemical processes over time can even be detected. The ultimate in microscopy is the electron microscope—to be discussed later. Research is being conducted into the development of new prototype microscopes that can become commercially available, providing better diagnostic and research capacities.



A confocal microscope provides three-dimensional images using pinholes and the extended depth of focus as described by wave optics. The right pinhole illuminates a tiny region of the sample in the focal plane. In-focus light rays from this tiny region pass through the dichroic mirror and the second pinhole to a detector and a computer. Out-of-focus light rays are blocked. The pinhole is scanned sideways to form an image of the entire focal plane. The pinhole can then be scanned up and down to gather images from different focal planes. The result is a three-dimensional image of the specimen.

Section Summary

- To improve microscope images, various techniques utilizing the wave characteristics of light have been developed. Many of these enhance contrast with interference effects.

Conceptual Questions

Exercise:

Problem:

Explain how microscopes can use wave optics to improve contrast and why this is important.

Exercise:

Problem:

A bright white light under water is collimated and directed upon a prism. What range of colors does one see emerging?

Glossary

confocal microscopes
microscopes that use the extended focal region to obtain three-dimensional images rather than two-dimensional images

contrast
the difference in intensity between objects and the background on which they are observed

interference microscopes
microscopes that enhance contrast between objects and background by superimposing a reference beam of light upon the light emerging from the sample

phase-contrast microscope
microscope utilizing wave interference and differences in phases to enhance contrast

polarization microscope
microscope that enhances contrast by utilizing a wave characteristic of light, useful for objects that are optically active

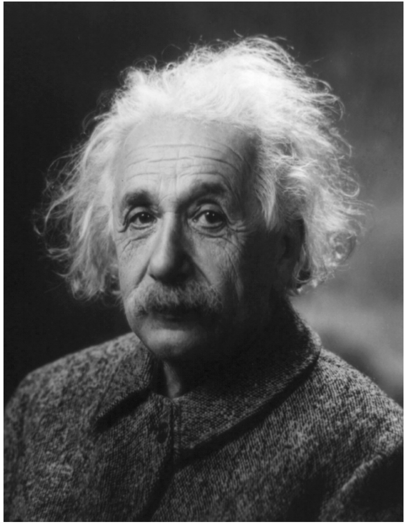
ultraviolet (UV) microscopes
microscopes constructed with special lenses that transmit UV rays and utilize photographic or electronic techniques to record images

Introduction to Special Relativity class="introduction"

Special relativity explains why traveling to other star systems, such as these in the Orion Nebula, is unreasonable using our current level of technology. (credit: s58y, Flickr)

Have you ever looked up at the night sky and dreamed of traveling to other planets in faraway star systems? Would there be other life forms? What would other worlds look like? You might imagine that such an amazing trip would be possible if we could just travel fast enough, but you will read in this chapter why this is not true. In 1905 Albert Einstein developed the theory of special relativity. This theory explains the limit on an object’s speed and describes the consequences.

Relativity. The word *relativity* might conjure an image of Einstein, but the idea did not begin with him. People have been exploring relativity for many centuries. Relativity is the study of how different observers measure the same event. Galileo and Newton developed the first correct version of classical relativity. Einstein developed the modern theory of relativity. Modern relativity is divided into two parts. *Special relativity* deals with observers who are moving at constant velocity. *General relativity* deals with observers who are undergoing acceleration. Einstein is famous because his theories of relativity made revolutionary predictions. Most importantly, his theories have been verified to great precision in a vast range of experiments, altering forever our concept of space and time.

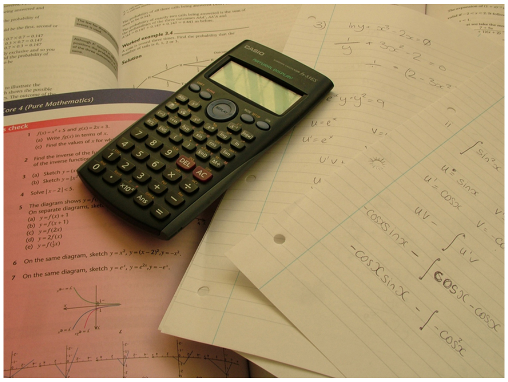


Many people think that Albert Einstein (1879–1955) was the greatest physicist of the 20th century. Not only did he develop modern relativity, thus revolutionizing our concept of the universe, he also made fundamental contributions to the foundations of quantum mechanics. (credit: The Library of Congress)

It is important to note that although classical mechanics, in general, and classical relativity, in particular, are limited, they are extremely good approximations for large, slow-moving objects. Otherwise, we could not use classical physics to launch satellites or build bridges. In the classical limit (objects larger than submicroscopic and moving slower than about 1% of the speed of light), relativistic mechanics becomes the same as classical mechanics. This fact will be noted at appropriate places throughout this chapter.

Einstein’s Postulates

- State and explain both of Einstein’s postulates.
- Explain what an inertial frame of reference is.
- Describe one way the speed of light can be changed.



Special relativity resembles trigonometry in that both are reliable because they are based on postulates that flow one from another in a logical way. (credit: Jon Oakley, Flickr)

Have you ever used the Pythagorean Theorem and gotten a wrong answer? Probably not, unless you made a mistake in either your algebra or your arithmetic. Each time you perform the same calculation, you know that the answer will be the same. Trigonometry is reliable because of the certainty that one part always flows from another in a logical way. Each part is based on a set of postulates, and you can always connect the parts by applying those postulates. Physics is the same way with the exception that *all* parts must describe nature. If we are careful to choose the correct postulates, then our theory will follow and will be verified by experiment.

Einstein essentially did the theoretical aspect of this method for **relativity**. With two deceptively simple postulates and a careful consideration of how measurements are made, he produced the theory of **special relativity.**

Einstein’s First Postulate

The first postulate upon which Einstein based the theory of special relativity relates to reference frames. All velocities are measured relative to some frame of reference. For example, a car’s motion is measured relative to its starting point or the road it is moving over, a projectile’s motion is measured relative to the surface it was launched from, and a planet’s orbit is measured relative to the star it is orbiting around. The simplest frames of reference are those that are not accelerated and are not rotating. Newton’s first law, the law of inertia, holds exactly in such a frame.

Note:
Inertial Reference Frame
An **inertial frame of reference** is a reference frame in which a body at rest remains at rest and a body in motion moves at a constant speed in a straight line unless acted on by an outside force.

The laws of physics seem to be simplest in inertial frames. For example, when you are in a plane flying at a constant altitude and speed, physics seems to work exactly the same as if you were standing on the surface of the Earth. However, in a plane that is taking off, matters are somewhat more complicated. In these cases, the net force on an object, F, is not equal to the product of mass and acceleration, ma. Instead, F is equal to ma plus a fictitious force. This situation is not as simple as in an inertial frame. Not only are laws of physics simplest in inertial frames, but they should be the same in all inertial frames, since there is no preferred frame and no absolute motion. Einstein incorporated these ideas into his **first postulate of special relativity**.

Note:
First Postulate of Special Relativity
The laws of physics are the same and can be stated in their simplest form in all inertial frames of reference.

As with many fundamental statements, there is more to this postulate than meets the eye. The laws of physics include only those that satisfy this postulate. We shall find that the definitions of relativistic momentum and energy must be altered to fit. Another outcome of this postulate is the famous equation $E = mc^2$.

Einstein's Second Postulate

The second postulate upon which Einstein based his theory of special relativity deals with the speed of light. Late in the 19th century, the major tenets of classical physics were well established. Two of the most important were the laws of electricity and magnetism and Newton's laws. In particular, the laws of electricity and magnetism predict that light travels at $c = 3.00 \times 10^8 \text{ m/s}$ in a vacuum, but they do not specify the frame of reference in which light has this speed.

There was a contradiction between this prediction and Newton's laws, in which velocities add like simple vectors. If the latter were true, then two observers moving at different speeds would see light traveling at different speeds. Imagine what a light wave would look like to a person traveling along with it at a speed c. If such a motion were possible then the wave would be stationary relative to the observer. It would have electric and magnetic fields that varied in strength at various distances from the observer but were constant in time. This is not allowed by Maxwell's equations. So either Maxwell's equations are wrong, or an object with mass cannot travel at speed c. Einstein concluded that the latter is true. An object with mass cannot travel at speed c. This conclusion implies that light in a vacuum must always travel at speed c relative to any observer. Maxwell's equations are correct, and Newton's addition of velocities is not correct for light.

Investigations such as Young’s double slit experiment in the early-1800s had convincingly demonstrated that light is a wave. Many types of waves were known, and all travelled in some medium. Scientists therefore assumed that a medium carried light, even in a vacuum, and light travelled at a speed *c* relative to that medium. Starting in the mid-1880s, the American physicist A. A. Michelson, later aided by E. W. Morley, made a series of direct measurements of the speed of light. The results of their measurements were startling.

Note:
Michelson-Morley Experiment
The **Michelson-Morley experiment** demonstrated that the speed of light in a vacuum is independent of the motion of the Earth about the Sun.

The eventual conclusion derived from this result is that light, unlike mechanical waves such as sound, does not need a medium to carry it. Furthermore, the Michelson-Morley results implied that the speed of light *c* is independent of the motion of the source relative to the observer. That is, everyone observes light to move at speed *c* regardless of how they move relative to the source or one another. For a number of years, many scientists tried unsuccessfully to explain these results and still retain the general applicability of Newton’s laws.

It was not until 1905, when Einstein published his first paper on special relativity, that the currently accepted conclusion was reached. Based mostly on his analysis that the laws of electricity and magnetism would not allow another speed for light, and only slightly aware of the Michelson-Morley experiment, Einstein detailed his **second postulate of special relativity**.

Note:
Second Postulate of Special Relativity

The speed of light c is a constant, independent of the relative motion of the source.

Deceptively simple and counterintuitive, this and the first postulate leave all else open for change. Some fundamental concepts do change. Among the changes are the loss of agreement on the elapsed time for an event, the variation of distance with speed, and the realization that matter and energy can be converted into one another. You will read about these concepts in the following sections.

Note:
Misconception Alert: Constancy of the Speed of Light
The speed of light is a constant $c = 3.00 \times 10^8\ \text{m/s}$ *in a vacuum*. If you remember the effect of the index of refraction from The Law of Refraction, the speed of light is lower in matter.

Exercise:
Check Your Understanding

Problem: Explain how special relativity differs from general relativity.

Solution:
Answer

Special relativity applies only to unaccelerated motion, but general relativity applies to accelerated motion.

Section Summary

- Relativity is the study of how different observers measure the same event.

- Modern relativity is divided into two parts. Special relativity deals with observers who are in uniform (unaccelerated) motion, whereas general relativity includes accelerated relative motion and gravity. Modern relativity is correct in all circumstances and, in the limit of low velocity and weak gravitation, gives the same predictions as classical relativity.
- An inertial frame of reference is a reference frame in which a body at rest remains at rest and a body in motion moves at a constant speed in a straight line unless acted on by an outside force.
- Modern relativity is based on Einstein’s two postulates. The first postulate of special relativity is the idea that the laws of physics are the same and can be stated in their simplest form in all inertial frames of reference. The second postulate of special relativity is the idea that the speed of light c is a constant, independent of the relative motion of the source.
- The Michelson-Morley experiment demonstrated that the speed of light in a vacuum is independent of the motion of the Earth about the Sun.

Conceptual Questions

Exercise:

Problem:

Which of Einstein’s postulates of special relativity includes a concept that does not fit with the ideas of classical physics? Explain.

Exercise:

Problem:

Is Earth an inertial frame of reference? Is the Sun? Justify your response.

Exercise:

Problem:

When you are flying in a commercial jet, it may appear to you that the airplane is stationary and the Earth is moving beneath you. Is this point of view valid? Discuss briefly.

Glossary

relativity
: the study of how different observers measure the same event

special relativity
: the theory that, in an inertial frame of reference, the motion of an object is relative to the frame from which it is viewed or measured

inertial frame of reference
: a reference frame in which a body at rest remains at rest and a body in motion moves at a constant speed in a straight line unless acted on by an outside force

first postulate of special relativity
: the idea that the laws of physics are the same and can be stated in their simplest form in all inertial frames of reference

second postulate of special relativity
: the idea that the speed of light c is a constant, independent of the source

Michelson-Morley experiment
: an investigation performed in 1887 that proved that the speed of light in a vacuum is the same in all frames of reference from which it is viewed

Simultaneity And Time Dilation

- Describe simultaneity.
- Describe time dilation.
- Calculate γ.
- Compare proper time and the observer’s measured time.
- Explain why the twin paradox is a false paradox.

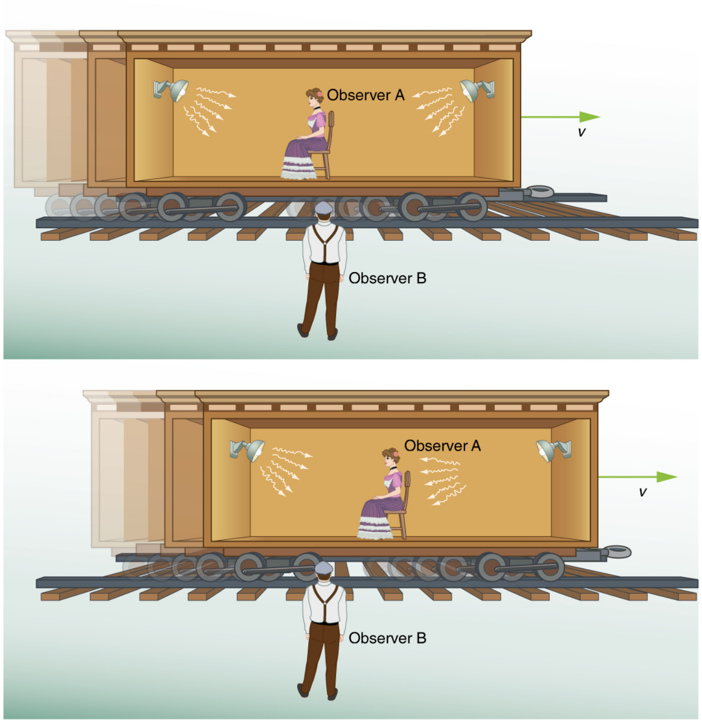
Elapsed time for a foot race is the same for all observers, but at relativistic speeds, elapsed time depends on the relative motion of the observer and the event that is observed. (credit: Jason Edward Scott Bain, Flickr)

Do time intervals depend on who observes them? Intuitively, we expect the time for a process, such as the elapsed time for a foot race, to be the same for all observers. Our experience has been that disagreements over elapsed time have to do with the accuracy of measuring time. When we carefully consider just how time is measured, however, we will find that elapsed time depends on the relative motion of an observer with respect to the process being measured.

Simultaneity

Consider how we measure elapsed time. If we use a stopwatch, for example, how do we know when to start and stop the watch? One method is to use the arrival of light from the event, such as observing a light turning green to start a drag race. The timing will be more accurate if some sort of electronic detection is used, avoiding human reaction times and other complications.

Now suppose we use this method to measure the time interval between two flashes of light produced by flash lamps. (See [link].) Two flash lamps with observer A midway between them are on a rail car that moves to the right relative to observer B. Observer B arranges for the light flashes to be emitted just as A passes B, so that both A and B are equidistant from the lamps when the light is emitted. Observer B measures the time interval between the arrival of the light flashes. According to postulate 2, the speed of light is not affected by the motion of the lamps relative to B. Therefore, light travels equal distances to him at equal speeds. Thus observer B measures the flashes to be simultaneous.



Observer B measures the elapsed time between the arrival of light flashes as described in the text. Observer A moves with the lamps on a rail car. Observer B perceives that the light flashes occurred simultaneously. Observer A perceives that the light on the right flashes before the light on the left.

Now consider what observer B sees happen to observer A. Observer B perceives light from the right reaching observer A before light from the left, because she has moved towards that flash lamp, lessening the distance the light must travel and reducing the time it takes to get to her. Light travels at speed *c* relative to both observers, but observer B remains equidistant between the points where the flashes were emitted, while A gets closer to the emission point on the right. From observer B’s point of view, then, there is a time interval between the arrival of the flashes to observer A. In observer A's frame of reference, the flashes occur at different times. Observer B measures the flashes to arrive simultaneously relative to him but not relative to A.

Now consider what observer A sees happening. She sees the light from the right arriving before light from the left. Since both lamps are the same distance from her in her reference frame, from her perspective, the right flash occurred before the left flash. Here a relative velocity between observers affects whether two events are observed to be simultaneous. *Simultaneity is not absolute*

This illustrates the power of clear thinking. We might have guessed incorrectly that if light is emitted simultaneously, then two observers halfway between the sources would see the flashes simultaneously. But careful analysis shows this not to be the case. Einstein was brilliant at this type of *thought experiment* (in German, “Gedankenexperiment”). He very carefully considered how an observation is made and disregarded what might seem obvious. The validity of thought experiments, of course, is determined by actual observation. The genius of Einstein is evidenced by the fact that experiments have repeatedly confirmed his theory of relativity.

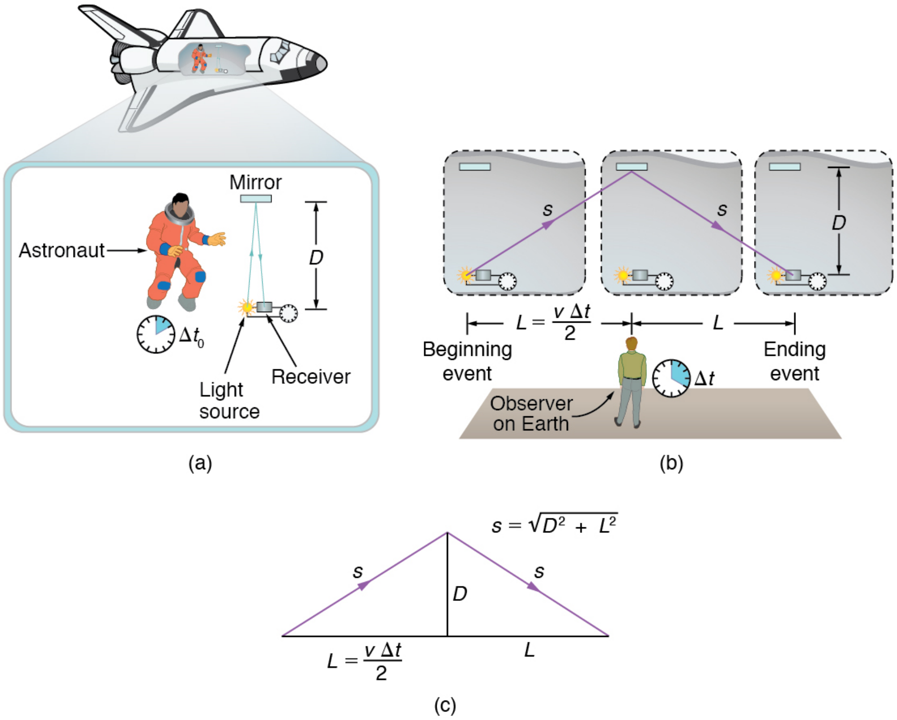
In summary: Two events are defined to be simultaneous if an observer measures them as occurring at the same time (such as by receiving light from the events). Two events are not necessarily simultaneous to all observers.

Time Dilation

The consideration of the measurement of elapsed time and simultaneity leads to an important relativistic effect.

Note:
Time dilation
Time dilation is the phenomenon of time passing slower for an observer who is moving relative to another observer.

Suppose, for example, an astronaut measures the time it takes for light to cross her ship, bounce off a mirror, and return. (See [link].) How does the elapsed time the astronaut measures compare with the elapsed time measured for the same event by a person on the Earth? Asking this question (another thought experiment) produces a profound result. We find that the elapsed time for a process depends on who is measuring it. In this case, the time measured by the astronaut is smaller than the time measured by the Earth-bound observer. The passage of time is different for the observers because the distance the light travels in the astronaut’s frame is smaller than in the Earth-bound frame. Light travels at the same speed in each frame, and so it will take longer to travel the greater distance in the Earth-bound frame.



(a) An astronaut measures the time Δt_0 for light to cross her ship using an electronic timer. Light travels a distance $2D$ in the astronaut's frame. (b) A person on the Earth sees the light follow the longer path $2s$ and take a longer time Δt. (c) These triangles are used to find the relationship between the two distances $2D$ and $2s$.

To quantitatively verify that time depends on the observer, consider the paths followed by light as seen by each observer. (See [link](c).) The astronaut sees the light travel straight across and back for a total distance of $2D$, twice the width of her ship. The Earth-bound observer sees the light travel a total distance $2s$. Since the ship is moving at speed v to the right relative to the Earth, light moving to the right hits the mirror in this frame. Light travels at a speed c in both frames, and because time is the distance divided by speed, the time measured by the astronaut is

Equation:

$$\Delta t_0 = \frac{2D}{c}.$$

This time has a separate name to distinguish it from the time measured by the Earth-bound observer.

Note:
Proper Time
Proper time Δt_0 is the time measured by an observer at rest relative to the event being observed.

In the case of the astronaut observe the reflecting light, the astronaut measures proper time. The time measured by the Earth-bound observer is
Equation:

$$\Delta t = \frac{2s}{c}.$$

To find the relationship between Δt_0 and Δt, consider the triangles formed by D and s. (See [link](c).) The third side of these similar triangles is L, the distance the astronaut moves as the light goes across her ship. In the frame of the Earth-bound observer,
Equation:

$$L = \frac{v\Delta t}{2}.$$

Using the Pythagorean Theorem, the distance s is found to be
Equation:

$$s = \sqrt{D^2 + \left(\frac{v\Delta t}{2}\right)^2}.$$

Substituting s into the expression for the time interval Δt gives

Equation:

$$\Delta t = \frac{2s}{c} = \frac{2\sqrt{D^2 + \left(\frac{v\Delta t}{2}\right)^2}}{c}.$$

We square this equation, which yields

Equation:

$$(\Delta t)^2 = \frac{4\left(D^2 + \frac{v^2(\Delta t)^2}{4}\right)}{c^2} = \frac{4D^2}{c^2} + \frac{v^2}{c^2}(\Delta t)^2.$$

Note that if we square the first expression we had for Δt_0, we get $(\Delta t_0)^2 = \frac{4D^2}{c^2}$. This term appears in the preceding equation, giving us a means to relate the two time intervals. Thus,

Equation:

$$(\Delta t)^2 = (\Delta t_0)^2 + \frac{v^2}{c^2}(\Delta t)^2.$$

Gathering terms, we solve for Δt:

Equation:

$$(\Delta t)^2\left(1 - \frac{v^2}{c^2}\right) = (\Delta t_0)^2.$$

Thus,

Equation:

$$(\Delta t)^2 = \frac{(\Delta t_0)^2}{1 - \frac{v^2}{c^2}}.$$

Taking the square root yields an important relationship between elapsed times:

Equation:

$$\Delta t = \frac{\Delta t_0}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma \Delta t_0,$$

where

Equation:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

This equation for Δt is truly remarkable. First, as contended, elapsed time is not the same for different observers moving relative to one another, even though both are in inertial frames. Proper time Δt_0 measured by an observer, like the astronaut moving with the apparatus, is smaller than time measured by other observers. Since those other observers measure a longer time Δt, the effect is called time dilation. The Earth-bound observer sees time dilate (get longer) for a system moving relative to the Earth. Alternatively, according to the Earth-bound observer, time slows in the moving frame, since less time passes there. All clocks moving relative to an observer, including biological clocks such as aging, are observed to run slow compared with a clock stationary relative to the observer.

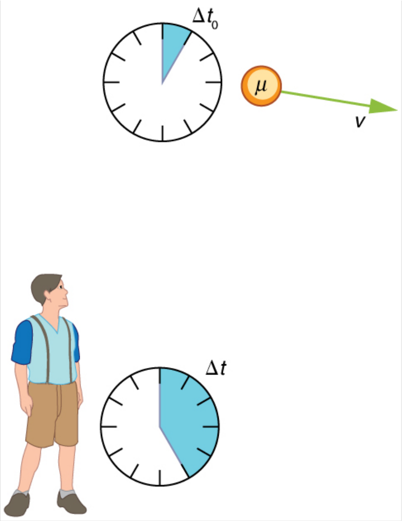
Note that if the relative velocity is much less than the speed of light ($v << c$), then $\frac{v^2}{c^2}$ is extremely small, and the elapsed times Δt and Δt_0 are nearly equal. At low velocities, modern relativity approaches classical physics—our everyday experiences have very small relativistic effects.

The equation $\Delta t = \gamma \Delta t_0$ also implies that relative velocity cannot exceed the speed of light. As v approaches c, Δt approaches infinity. This would imply that time in the astronaut’s frame stops at the speed of light. If v exceeded c, then we would be taking the square root of a negative number, producing an imaginary value for Δt.

There is considerable experimental evidence that the equation $\Delta t = \gamma \Delta t_0$ is correct. One example is found in cosmic ray particles that continuously rain down on the Earth from deep space. Some collisions of these particles with nuclei in the upper atmosphere result in short-lived particles called muons. The half-life (amount of time for half of a material to decay) of a muon is $1.52\ \mu\text{s}$ when it is at rest relative to the observer who measures the half-life. This is the proper time Δt_0. Muons produced by cosmic ray particles have a range of velocities, with some moving near the speed of light. It has been found that the muon’s half-life as measured by an Earth-bound observer (Δt) varies with velocity exactly as predicted by the equation $\Delta t = \gamma \Delta t_0$. The faster the muon moves, the longer it lives. We on the Earth see the muon’s half-life time dilated—as viewed from our frame, the muon decays more slowly than it does when at rest relative to us.

Example:
Calculating Δt for a Relativistic Event: How Long Does a Speedy Muon Live?

Suppose a cosmic ray colliding with a nucleus in the Earth’s upper atmosphere produces a muon that has a velocity $v = 0.950c$. The muon then travels at constant velocity and lives $1.52\ \mu\text{s}$ as measured in the muon’s frame of reference. (You can imagine this as the muon’s internal clock.) How long does the muon live as measured by an Earth-bound observer? (See [link].)



A muon in the Earth's atmosphere lives longer as measured by an Earth-bound observer than measured by the muon's internal clock.

Strategy
A clock moving with the system being measured observes the proper time, so the time we are given is $\Delta t_0 = 1.52\ \mu$s. The Earth-bound observer measures Δt as given by the equation $\Delta t = \gamma \Delta t_0$. Since we know the velocity, the calculation is straightforward.

Solution
1) Identify the knowns. $v = 0.950c$, $\Delta t_0 = 1.52\ \mu$s
2) Identify the unknown. Δt
3) Choose the appropriate equation.
Use,

Equation:

$$\Delta t = \gamma \Delta t_0,$$

where

Equation:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

4) Plug the knowns into the equation.

First find γ.

Equation:

$$\begin{aligned}\gamma &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\\ &= \frac{1}{\sqrt{1-\frac{(0.950c)^2}{c^2}}}\\ &= \frac{1}{\sqrt{1-(0.950)^2}}\\ &= 3.20.\end{aligned}$$

Use the calculated value of γ to determine Δt.

Equation:

$$\begin{aligned}\Delta t &= \gamma \Delta t_0\\ &= (3.20)(1.52\ \mu\text{s})\\ &= 4.87\ \mu\text{s}\end{aligned}$$

Discussion

One implication of this example is that since $\gamma = 3.20$ at 95.0% of the speed of light ($v = 0.950c$), the relativistic effects are significant. The two time intervals differ by this factor of 3.20, where classically they would be the same. Something moving at $0.950c$ is said to be highly relativistic.

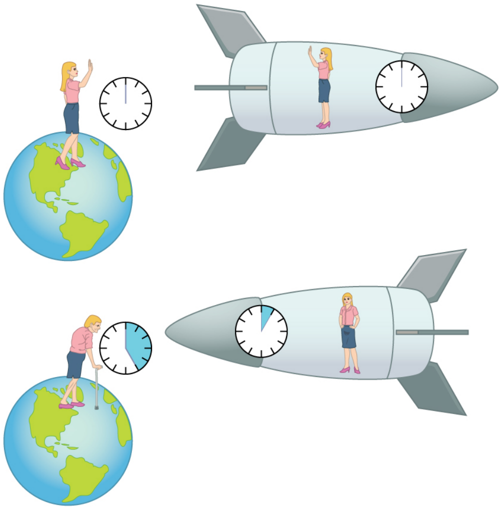
Another implication of the preceding example is that everything an astronaut does when moving at 95.0% of the speed of light relative to the Earth takes 3.20 times longer when observed from the Earth. Does the astronaut sense this? Only if she looks outside her spaceship. All methods of measuring time in her frame will be affected by the same factor of 3.20. This includes her wristwatch, heart rate, cell metabolism rate, nerve impulse rate, and so on. She will have no way of telling, since all of her clocks will agree with one another because their relative velocities are zero. Motion is relative, not absolute. But what if she does look out the window?

Note:
Real-World Connections
It may seem that special relativity has little effect on your life, but it is probably more important than you realize. One of the most common effects is through the Global Positioning System (GPS). Emergency vehicles, package delivery services, electronic maps, and communications devices are just a few of the common uses of GPS, and the GPS system could not work without taking into account relativistic effects. GPS satellites rely on precise time measurements to communicate. The signals travel at relativistic speeds. Without corrections for time dilation, the satellites could not communicate, and the GPS system would fail within minutes.

The Twin Paradox

An intriguing consequence of time dilation is that a space traveler moving at a high velocity relative to the Earth would age less than her Earth-bound twin. Imagine the astronaut moving at such a velocity that $\gamma = 30.0$, as in [link]. A trip that takes 2.00 years in her frame would take 60.0 years in her Earth-bound twin's frame. Suppose the astronaut traveled 1.00 year to another star system. She briefly explored the area, and then traveled 1.00 year back. If the astronaut was 40 years old when she left, she would be 42 upon her return. Everything on the Earth, however, would have aged 60.0 years. Her twin, if still alive, would be 100 years old.

The situation would seem different to the astronaut. Because motion is relative, the spaceship would seem to be stationary and the Earth would appear to move. (This is the sensation you have when flying in a jet.) If the astronaut looks out the window of the spaceship, she will see time slow down on the Earth by a factor of $\gamma = 30.0$. To her, the Earth-bound sister will have aged only 2/30 (1/15) of a year, while she aged 2.00 years. The two sisters cannot both be correct.



The twin paradox asks why the traveling twin ages less than the Earth-bound twin. That is the prediction we obtain if we consider the Earth-bound twin's frame. In the astronaut's frame, however, the Earth is moving and time runs slower there. Who is correct?

As with all paradoxes, the premise is faulty and leads to contradictory conclusions. In fact, the astronaut's motion is significantly different from that of the Earth-bound twin. The astronaut accelerates to a high velocity and then decelerates to view the star system. To return to the Earth, she again accelerates and decelerates. The Earth-bound twin does not experience these accelerations. So the situation is not symmetric, and it is not correct to claim that the astronaut will observe the same effects as her Earth-bound twin. If you use special relativity to examine the twin paradox, you must keep in mind that the theory is expressly based on inertial frames, which by definition are not accelerated or rotating. Einstein developed general relativity to deal with accelerated frames and with gravity, a prime source of acceleration. You can also use general relativity to address the twin paradox and, according to general relativity, the astronaut will age less. Some important conceptual aspects of general relativity are discussed in General Relativity and Quantum Gravity of this course.

In 1971, American physicists Joseph Hafele and Richard Keating verified time dilation at low relative velocities by flying extremely accurate atomic clocks around the Earth on commercial aircraft. They measured elapsed time to an accuracy of a few nanoseconds and compared it with the time measured by clocks left behind. Hafele and Keating's results were within experimental uncertainties of the predictions of relativity. Both special and general relativity had to be taken into account, since gravity and accelerations were involved as well as relative motion.

Exercise:

Check Your Understanding

Problem:1. What is γ if $v = 0.650c$?

Solution

$$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}} = \frac{1}{\sqrt{1-\frac{(0.650c)^2}{c^2}}} = 1.32$$

2. A particle travels at 1.90×10^8 m/s and lives 2.10×10^{-8} s when at rest relative to an observer. How long does the particle live as viewed in the laboratory?

Solution:

$$\Delta t = \frac{\Delta_t^0}{\sqrt{1 - \frac{v^2}{c^2}}} = \frac{2.10 \times 10^{-8}\ \text{s}}{\sqrt{1 - \frac{(1.90 \times 10^8\ \text{m/s})^2}{(3.00 \times 10^8\ \text{m/s})^2}}} = 2.71 \times 10^{-8}\ \text{s}$$

Section Summary

- Two events are defined to be simultaneous if an observer measures them as occurring at the same time. They are not necessarily simultaneous to all observers—simultaneity is not absolute.
- Time dilation is the phenomenon of time passing slower for an observer who is moving relative to another observer.
- Observers moving at a relative velocity v do not measure the same elapsed time for an event. Proper time Δt_0 is the time measured by an observer at rest relative to the event being observed. Proper time is related to the time Δt measured by an Earth-bound observer by the equation
  Equation:

  $$\Delta t = \frac{\Delta t_0}{\sqrt{1 - \frac{v^2}{c^2}}} = \gamma \Delta t_0,$$

  where
  Equation:

  $$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}.$$

- The equation relating proper time and time measured by an Earth-bound observer implies that relative velocity cannot exceed the speed of light.
- The twin paradox asks why a twin traveling at a relativistic speed away and then back towards the Earth ages less than the Earth-bound twin. The premise to the paradox is faulty because the traveling twin is accelerating. Special relativity does not apply to accelerating frames of reference.

- Time dilation is usually negligible at low relative velocities, but it does occur, and it has been verified by experiment.

Conceptual Questions

Exercise:

Problem:

Does motion affect the rate of a clock as measured by an observer moving with it? Does motion affect how an observer moving relative to a clock measures its rate?

Exercise:

Problem:

To whom does the elapsed time for a process seem to be longer, an observer moving relative to the process or an observer moving with the process? Which observer measures proper time?

Exercise:

Problem:

How could you travel far into the future without aging significantly? Could this method also allow you to travel into the past?

Problems & Exercises

Exercise:

Problem: (a) What is γ if $v = 0.250c$? (b) If $v = 0.500c$?

Solution:

(a) 1.0328

(b) 1.15

Exercise:

Problem: (a) What is γ if $v = 0.100c$? (b) If $v = 0.900c$?

Exercise:

Problem:

Particles called π-mesons are produced by accelerator beams. If these particles travel at $2.70 \times 10^8\ \text{m/s}$ and live 2.60×10^{-8} s when at rest relative to an observer, how long do they live as viewed in the laboratory?

Solution:

5.96×10^{-8} s

Exercise:

Problem:

Suppose a particle called a kaon is created by cosmic radiation striking the atmosphere. It moves by you at $0.980c$, and it lives 1.24×10^{-8} s when at rest relative to an observer. How long does it live as you observe it?

Exercise:

Problem:

A neutral π-meson is a particle that can be created by accelerator beams. If one such particle lives 1.40×10^{-16} s as measured in the laboratory, and 0.840×10^{-16} s when at rest relative to an observer, what is its velocity relative to the laboratory?

Solution:

$0.800c$

Exercise:

Problem:

A neutron lives 900 s when at rest relative to an observer. How fast is the neutron moving relative to an observer who measures its life span to be 2065 s?

Exercise:

Problem:

If relativistic effects are to be less than 1%, then γ must be less than 1.01. At what relative velocity is $\gamma = 1.01$?

Solution:

$0.140c$

Exercise:

Problem:

If relativistic effects are to be less than 3%, then γ must be less than 1.03. At what relative velocity is $\gamma = 1.03$?

Exercise:

Problem:

(a) At what relative velocity is $\gamma = 1.50$? (b) At what relative velocity is $\gamma = 100$?

Solution:

(a) $0.745c$

(b) $0.99995c$ (to five digits to show effect)

Exercise:

Problem:

(a) At what relative velocity is $\gamma = 2.00$? (b) At what relative velocity is $\gamma = 10.0$?

Exercise:

Problem: **Unreasonable Results**

(a) Find the value of γ for the following situation. An Earth-bound observer measures 23.9 h to have passed while signals from a high-velocity space probe indicate that $24.0\ \text{h}$ have passed on board. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) 0.996

(b) γ cannot be less than 1.

(c) Assumption that time is longer in moving ship is unreasonable.

Glossary

time dilation
: the phenomenon of time passing slower to an observer who is moving relative to another observer

proper time
: Δt_0. the time measured by an observer at rest relative to the event being observed: $\Delta t = \frac{\Delta t_0}{\sqrt{1-\frac{v^2}{c^2}}} = \gamma \Delta t_0$, where $\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$

twin paradox
: this asks why a twin traveling at a relativistic speed away and then back towards the Earth ages less than the Earth-bound twin. The

premise to the paradox is faulty because the traveling twin is accelerating, and special relativity does not apply to accelerating frames of reference

Length Contraction

- Describe proper length.
- Calculate length contraction.
- Explain why we don’t notice these effects at everyday scales.

People might describe distances differently, but at relativistic speeds, the distances really are different. (credit: Corey Leopold, Flickr)

Have you ever driven on a road that seems like it goes on forever? If you look ahead, you might say you have about 10 km left to go. Another traveler might say the road ahead looks like it’s about 15 km long. If you both measured the road, however, you would agree. Traveling at everyday speeds, the distance you both measure would be the same. You will read in this section, however, that this is not true at relativistic speeds. Close to the speed of light, distances measured are not the same when measured by different observers.

Proper Length

One thing all observers agree upon is relative speed. Even though clocks measure different elapsed times for the same process, they still agree that relative speed, which is distance divided by elapsed time, is the same. This implies that distance, too, depends on the observer’s relative motion. If two observers see different times, then they must also see different distances for relative speed to be the same to each of them.

The muon discussed in [link] illustrates this concept. To an observer on the Earth, the muon travels at $0.950c$ for $7.05\ \mu\text{s}$ from the time it is produced until it decays. Thus it travels a distance

Equation:

$$L_0 = v\Delta t = (0.950)(3.00 \times 10^8\ \text{m/s})(7.05 \times 10^{-6}\ \text{s}) = 2.01\ \text{km}$$

relative to the Earth. In the muon’s frame of reference, its lifetime is only $2.20\ \mu\text{s}$. It has enough time to travel only

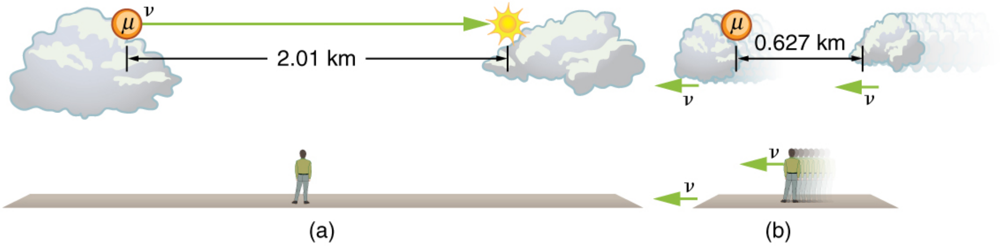
Equation:

$$L = v\Delta t_0 = (0.950)(3.00 \times 10^8\ \text{m/s})(2.20 \times 10^{-6}\ \text{s}) = 0.627\ \text{km.}$$

The distance between the same two events (production and decay of a muon) depends on who measures it and how they are moving relative to it.

Note:
Proper Length
Proper length L_0 is the distance between two points measured by an observer who is at rest relative to both of the points.

The Earth-bound observer measures the proper length L_0, because the points at which the muon is produced and decays are stationary relative to the Earth. To the muon, the Earth, air, and clouds are moving, and so the distance L it sees is not the proper length.



(a) The Earth-bound observer sees the muon travel 2.01 km between clouds. (b) The muon sees itself travel the same path, but only a distance of 0.627 km. The Earth, air, and clouds are moving relative to the muon in its frame, and all appear to have smaller lengths along the direction of travel.

Length Contraction

To develop an equation relating distances measured by different observers, we note that the velocity relative to the Earth-bound observer in our muon example is given by

Equation:

$$v=\frac{L_0}{\Delta t}.$$

The time relative to the Earth-bound observer is Δt, since the object being timed is moving relative to this observer. The velocity relative to the moving observer is given by

Equation:

$$v=\frac{L}{\Delta t_0}.$$

The moving observer travels with the muon and therefore observes the proper time Δt_0. The two velocities are identical; thus,

Equation:

$$\frac{L_0}{\Delta t} = \frac{L}{\Delta t_0}.$$

We know that $\Delta t = \gamma \Delta t_0$. Substituting this equation into the relationship above gives

Equation:

$$L = \frac{L_0}{\gamma}.$$

Substituting for γ gives an equation relating the distances measured by different observers.

Note:

Length Contraction

Length contraction L is the shortening of the measured length of an object moving relative to the observer's frame.

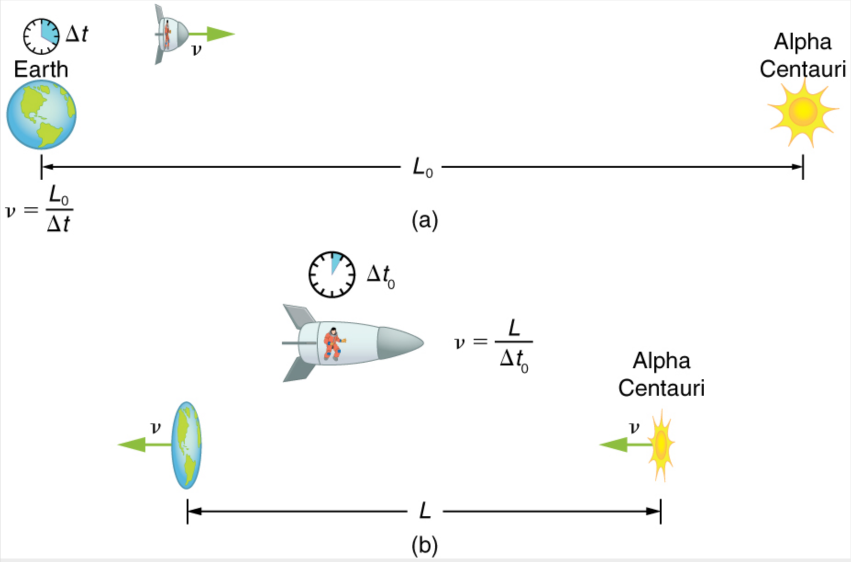
Equation:

$$L = L_0 \sqrt{1 - \frac{v^2}{c^2}}.$$

If we measure the length of anything moving relative to our frame, we find its length L to be smaller than the proper length L_0 that would be measured if the object were stationary. For example, in the muon's reference frame, the distance between the points where it was produced and where it decayed is shorter. Those points are fixed relative to the Earth but moving relative to the muon. Clouds and other objects are also contracted along the direction of motion in the muon's reference frame.

Example:
Calculating Length Contraction: The Distance between Stars Contracts when You Travel at High Velocity

Suppose an astronaut, such as the twin discussed in Simultaneity and Time Dilation, travels so fast that $\gamma = 30.00$. (a) She travels from the Earth to the nearest star system, Alpha Centauri, 4.300 light years (ly) away as measured by an Earth-bound observer. How far apart are the Earth and Alpha Centauri as measured by the astronaut? (b) In terms of c, what is her velocity relative to the Earth? You may neglect the motion of the Earth relative to the Sun. (See [link].)



(a) The Earth-bound observer measures the proper distance between the Earth and the Alpha Centauri. (b) The astronaut observes a length contraction, since the Earth and the Alpha Centauri move relative to her ship. She can travel this shorter distance in a smaller time (her proper time) without exceeding the speed of light.

Strategy
First note that a light year (ly) is a convenient unit of distance on an astronomical scale—it is the distance light travels in a year. For part (a), note that the 4.300 ly distance between the Alpha Centauri and the Earth is

the proper distance L_0, because it is measured by an Earth-bound observer to whom both stars are (approximately) stationary. To the astronaut, the Earth and the Alpha Centauri are moving by at the same velocity, and so the distance between them is the contracted length L. In part (b), we are given γ, and so we can find v by rearranging the definition of γ to express v in terms of c.

Solution for (a)

1. Identify the knowns. $L_0 - 4.300\text{ ly}; \gamma = 30.00$
2. Identify the unknown. L
3. Choose the appropriate equation. $L = \frac{L_0}{\gamma}$
4. Rearrange the equation to solve for the unknown.
   Equation:

$$\begin{aligned} L &= \frac{L_0}{\gamma} \\ &= \frac{4.300\text{ ly}}{30.00} \\ &= 0.1433\text{ ly} \end{aligned}$$

Solution for (b)

1. Identify the known. $\gamma = 30.00$
2. Identify the unknown. v in terms of c
3. Choose the appropriate equation. $\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$
4. Rearrange the equation to solve for the unknown.
   Equation:

$$\begin{aligned} \gamma &= \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \\ 30.00 &= \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \end{aligned}$$

   Squaring both sides of the equation and rearranging terms gives
   Equation:

$$900.0 = \frac{1}{1 - \frac{v^2}{c^2}}$$

so that
Equation:

$$1 - \frac{v^2}{c^2} = \frac{1}{900.0}$$

and
Equation:

$$\frac{v^2}{c^2} = 1 - \frac{1}{900.0} = 0.99888....$$

Taking the square root, we find
Equation:

$$\frac{v}{c} = 0.99944,$$

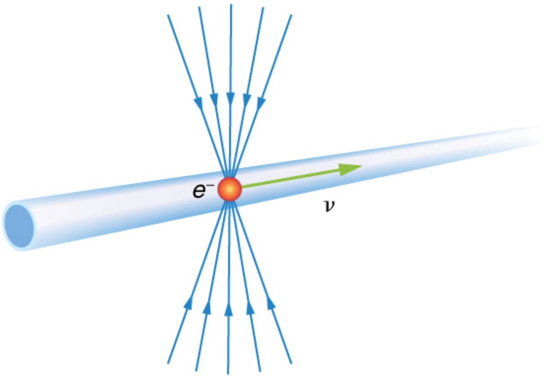
which is rearranged to produce a value for the velocity
Equation:

$$\text{v}= 0.9994c.$$

Discussion
First, remember that you should not round off calculations until the final result is obtained, or you could get erroneous results. This is especially true for special relativity calculations, where the differences might only be revealed after several decimal places. The relativistic effect is large here (γ=30.00), and we see that v is approaching (not equaling) the speed of light. Since the distance as measured by the astronaut is so much smaller, the astronaut can travel it in much less time in her frame.

People could be sent very large distances (thousands or even millions of light years) and age only a few years on the way if they traveled at extremely high velocities. But, like emigrants of centuries past, they would leave the Earth they know forever. Even if they returned, thousands to millions of years would have passed on the Earth, obliterating most of what now exists. There is also a more serious practical obstacle to traveling at such velocities; immensely greater energies than classical physics predicts would be needed to achieve such high velocities. This will be discussed in Relatavistic Energy.

Why don’t we notice length contraction in everyday life? The distance to the grocery shop does not seem to depend on whether we are moving or not. Examining the equation $L=L_0\sqrt{1-\frac{v^2}{c^2}}$, we see that at low velocities ($v<<c$) the lengths are nearly equal, the classical expectation. But length contraction is real, if not commonly experienced. For example, a charged particle, like an electron, traveling at relativistic velocity has electric field lines that are compressed along the direction of motion as seen by a stationary observer. (See [link].) As the electron passes a detector, such as a coil of wire, its field interacts much more briefly, an effect observed at particle accelerators such as the 3 km long Stanford Linear Accelerator (SLAC). In fact, to an electron traveling down the beam pipe at SLAC, the accelerator and the Earth are all moving by and are length contracted. The relativistic effect is so great than the accelerator is only 0.5 m long to the electron. It is actually easier to get the electron beam down the pipe, since the beam does not have to be as precisely aimed to get down a short pipe as it would down one 3 km long. This, again, is an experimental verification of the Special Theory of Relativity.



The electric field lines of a high-velocity charged particle are compressed along the direction of motion by length contraction. This produces a different signal when the particle goes through a coil, an experimentally verified effect of length contraction.

Exercise:
Check Your Understanding

Problem:

A particle is traveling through the Earth’s atmosphere at a speed of $0.750c$. To an Earth-bound observer, the distance it travels is 2.50 km. How far does the particle travel in the particle’s frame of reference?

Solution:
Answer
Equation:

$$\mathrm{L}{=}\mathrm{L}_0\sqrt{1-\frac{v^2}{c^2}}=(2.50\text{ km})\quad\overline{1-\frac{(0.750c)^2}{c^2}}=1.65\text{ km}$$

Summary

- All observers agree upon relative speed.
- Distance depends on an observer's motion. Proper length L_0 is the distance between two points measured by an observer who is at rest relative to both of the points. Earth-bound observers measure proper length when measuring the distance between two points that are stationary relative to the Earth.
- Length contraction L is the shortening of the measured length of an object moving relative to the observer's frame:
  Equation:

$$\mathrm{L}{=}\mathrm{L}_0\sqrt{1-\frac{v^2}{c^2}}=\frac{L_0}{\gamma}.$$

Conceptual Questions

Exercise:

Problem:

To whom does an object seem greater in length, an observer moving with the object or an observer moving relative to the object? Which observer measures the object's proper length?

Exercise:

Problem:

Relativistic effects such as time dilation and length contraction are present for cars and airplanes. Why do these effects seem strange to us?

Exercise:

Problem:

Suppose an astronaut is moving relative to the Earth at a significant fraction of the speed of light. (a) Does he observe the rate of his clocks to have slowed? (b) What change in the rate of Earth-bound clocks does he see? (c) Does his ship seem to him to shorten? (d) What about the distance between stars that lie on lines parallel to his motion? (e) Do he and an Earth-bound observer agree on his velocity relative to the Earth?

Problems & Exercises

Exercise:

Problem:

A spaceship, 200 m long as seen on board, moves by the Earth at $0.970c$. What is its length as measured by an Earth-bound observer?

Solution:

48.6 m

Exercise:

Problem:

How fast would a 6.0 m-long sports car have to be going past you in order for it to appear only 5.5 m long?

Exercise:

Problem:

(a) How far does the muon in [link] travel according to the Earth-bound observer? (b) How far does it travel as viewed by an observer moving with it? Base your calculation on its velocity relative to the Earth and the time it lives (proper time). (c) Verify that these two distances are related through length contraction $\gamma=3.20$.

Solution:

(a) 1.387 km = 1.39 km

(b) 0.433 km

(c) $$\begin{array}{rclcl} L & = & \frac{L_0}{\gamma} & = & \frac{1.387\times10^3\text{ m}}{3.20} \\ & = & 433.4\text{ m} & = & 0.433\text{ km} \end{array}$$

Thus, the distances in parts (a) and (b) are related when $\gamma = 3.20$.

Exercise:

Problem:

(a) How long would the muon in [link] have lived as observed on the Earth if its velocity was $0.0500c$? (b) How far would it have traveled as observed on the Earth? (c) What distance is this in the muon’s frame?

Exercise:

Problem:

(a) How long does it take the astronaut in [link] to travel 4.30 ly at $0.99944c$ (as measured by the Earth-bound observer)? (b) How long does it take according to the astronaut? (c) Verify that these two times are related through time dilation with $\gamma=30.00$ as given.

Solution:

(a) 4.303 y (to four digits to show any effect)

(b) 0.1434 y

(c) $\Delta t = \gamma \Delta t_0 \Rightarrow \gamma = \frac{\Delta t}{\Delta t_0} = \frac{4.303\text{ y}}{0.1434\text{ y}} = 30.0$

Thus, the two times are related when $\gamma \quad 30.00$.

Exercise:

Problem:

(a) How fast would an athlete need to be running for a 100-m race to look 100 yd long? (b) Is the answer consistent with the fact that relativistic effects are difficult to observe in ordinary circumstances? Explain.

Exercise:

Problem: Unreasonable Results

(a) Find the value of γ for the following situation. An astronaut measures the length of her spaceship to be 25.0 m, while an Earth-bound observer measures it to be 100 m. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) 0.250

(b) γ must be ≥1

(c) The Earth-bound observer must measure a shorter length, so it is unreasonable to assume a longer length.

Exercise:

Problem: Unreasonable Results

A spaceship is heading directly toward the Earth at a velocity of $0.800c$. The astronaut on board claims that he can send a canister toward the Earth at $1.20c$ relative to the Earth. (a) Calculate the velocity the canister must have relative to the spaceship. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

proper length

L_0; the distance between two points measured by an observer who is at rest relative to both of the points; Earth-bound observers measure proper length when measuring the distance between two points that are stationary relative to the Earth

length contraction

L, the shortening of the measured length of an object moving relative to the observer's frame: $\mathrm{L}=\mathrm{L}_0\sqrt{1-\frac{v^2}{c^2}}=\frac{L_0}{\gamma}$

Relativistic Addition of Velocities

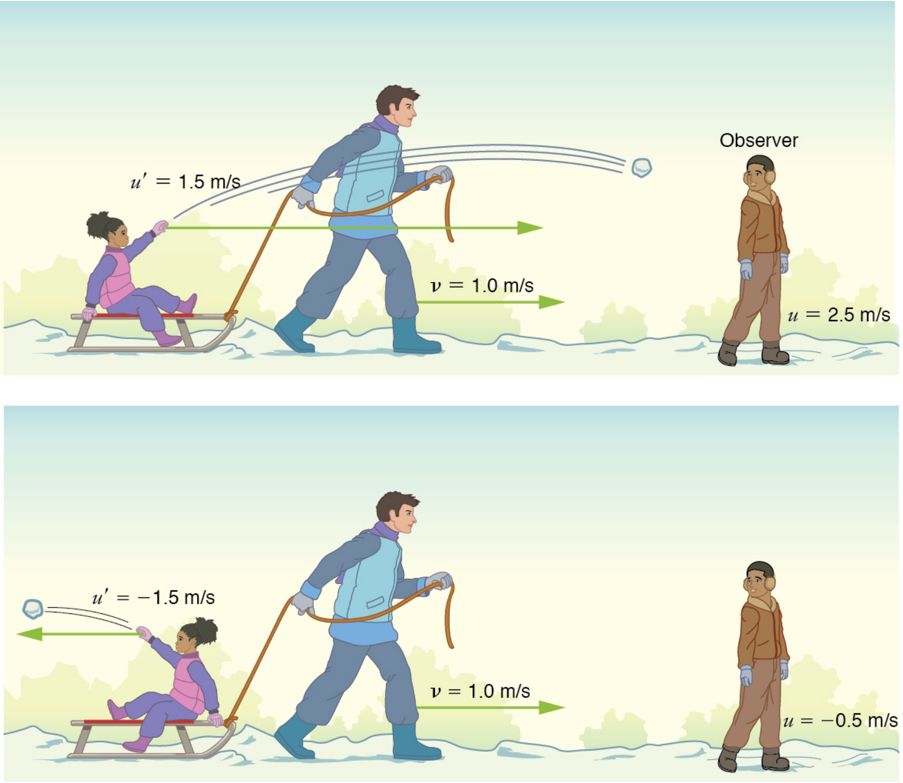
- Calculate relativistic velocity addition.
- Explain when relativistic velocity addition should be used instead of classical addition of velocities.
- Calculate relativistic Doppler shift.

The total velocity of a kayak, like this one on the Deerfield River in Massachusetts, is its velocity relative to the water as well as the water’s velocity relative to the riverbank. (credit: abkfenris, Flickr)

If you’ve ever seen a kayak move down a fast-moving river, you know that remaining in the same place would be hard. The river current pulls the kayak along. Pushing the oars back against the water can move the kayak forward in the water, but that only accounts for part of the velocity. The kayak’s motion is an example of classical addition of velocities. In classical physics, velocities add as vectors. The kayak’s velocity is the vector sum of its velocity relative to the water and the water’s velocity relative to the riverbank.

Classical Velocity Addition

For simplicity, we restrict our consideration of velocity addition to one-dimensional motion. Classically, velocities add like regular numbers in one-dimensional motion. (See [link].) Suppose, for example, a girl is riding in a sled at a speed 1.0 m/s relative to an observer. She throws a snowball first forward, then backward at a speed of 1.5 m/s relative to the sled. We denote direction with plus and minus signs in one dimension; in this example, forward is positive. Let v be the velocity of the sled relative to the Earth, u the velocity of the snowball relative to the Earth-bound observer, and $u\prime$ the velocity of the snowball relative to the sled.



Classically, velocities add like ordinary numbers in one-dimensional motion. Here the girl throws a snowball forward and then backward from a sled. The velocity of the sled relative to the Earth is $v=1.0\text{ m/s}$. The velocity of the snowball relative to the sled is $u\prime$, while its velocity relative to the Earth is u. Classically, $\text{u=v+u}\prime$.

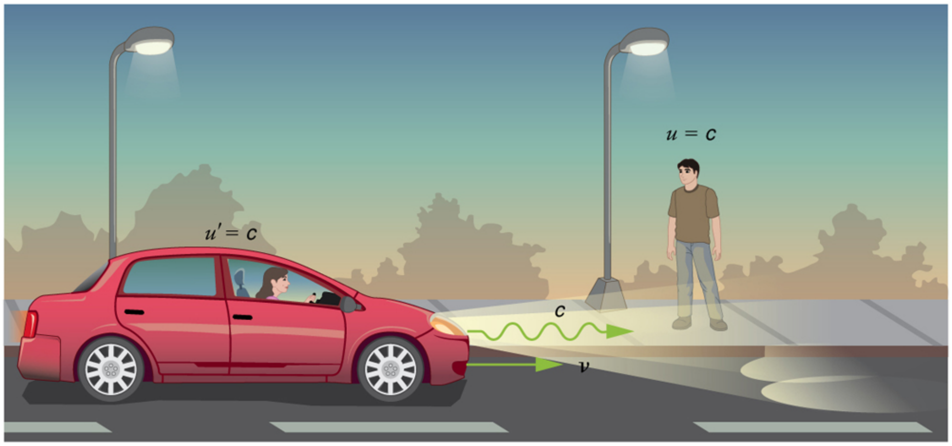
Note:
Classical Velocity Addition
Equation:

$$\text{u=v+u}\prime$$

Thus, when the girl throws the snowball forward, $u = 1.0\ \text{m/s} + 1.5\ \text{m/s} = 2.5\ \text{m/s}$. It makes good intuitive sense that the snowball will head towards the Earth-bound observer faster, because it is thrown forward from a moving vehicle. When the girl throws the snowball backward, $u = 1.0\ \text{m/s} + (-1.5\ \text{m/s}) = -0.5\ \text{m/s}$. The minus sign means the snowball moves away from the Earth-bound observer.

Relativistic Velocity Addition

The second postulate of relativity (verified by extensive experimental observation) says that classical velocity addition does not apply to light. Imagine a car traveling at night along a straight road, as in [link]. If classical velocity addition applied to light, then the light from the car's headlights would approach the observer on the sidewalk at a speed u=v+c. But we know that light will move away from the car at speed c relative to the driver of the car, and light will move towards the observer on the sidewalk at speed c, too.



According to experiment and the second postulate of relativity, light from the car’s headlights moves away from the car at speed c and towards the observer on the sidewalk at speed c. Classical velocity addition is not valid.

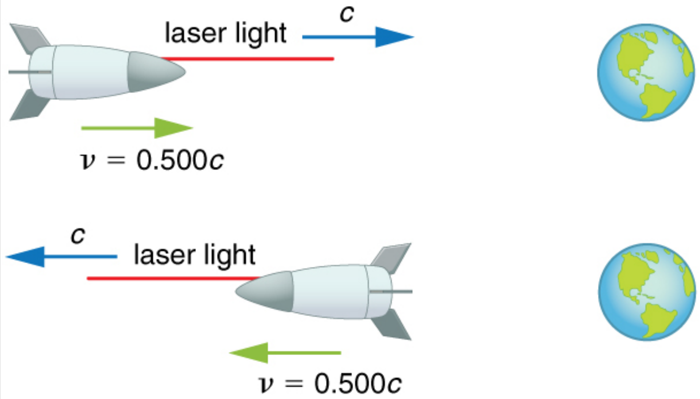
Note:
Relativistic Velocity Addition
Either light is an exception, or the classical velocity addition formula only works at low velocities. The latter is the case. The correct formula for one-dimensional **relativistic velocity addition** is
Equation:

$$u = \frac{v+u'}{1+\frac{vu'}{c^2}},$$

where v is the relative velocity between two observers, u is the velocity of an object relative to one observer, and u' is the velocity relative to the other observer. (For ease of visualization, we often choose to measure u in our reference frame, while someone moving at v relative to us measures u'.) Note that the term $\frac{vu'}{c^2}$ becomes very small at low velocities, and $u = \frac{v+u'}{1+\frac{vu'}{c^2}}$ gives a result very close to classical velocity addition. As before, we see that classical velocity addition is an excellent approximation to the correct relativistic formula for small velocities. No wonder that it seems correct in our experience.

Example:
Showing that the Speed of Light towards an Observer is Constant (in a Vacuum): The Speed of Light is the Speed of Light

Suppose a spaceship heading directly towards the Earth at half the speed of light sends a signal to us on a laser-produced beam of light. Given that the light leaves the ship at speed c as observed from the ship, calculate the speed at which it approaches the Earth.



Strategy

Because the light and the spaceship are moving at relativistic speeds, we cannot use simple velocity addition. Instead, we can determine the speed at which the light approaches the Earth using relativistic velocity addition.

Solution

1. Identify the knowns. v$=0.500c$; $u\prime = c$
2. Identify the unknown. u
3. Choose the appropriate equation. $u = \frac{v+u\prime}{1+\frac{vu\prime}{c^2}}$
4. Plug the knowns into the equation.
   Equation:

$$
\begin{aligned}
u &= \frac{v+u'}{1+\frac{vu'}{c^2}} \\
&= \frac{0.500c+c}{1+\frac{(0.500c)(c)}{c^2}} \\
&= \frac{(0.500+1)c}{1+\frac{0.500c^2}{c^2}} \\
&= \frac{1.500c}{1+0.500} \\
&= \frac{1.500c}{1.500} \\
&= c
\end{aligned}
$$

Discussion

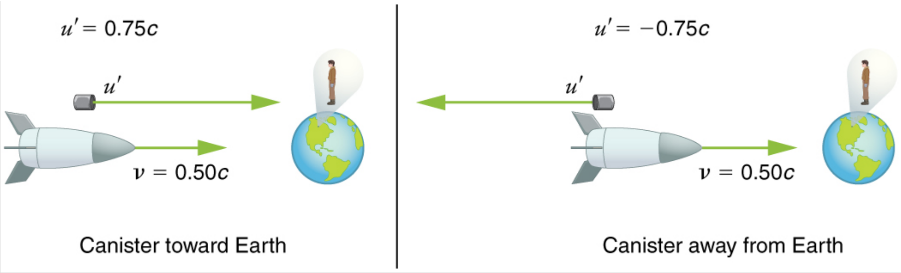
Relativistic velocity addition gives the correct result. Light leaves the ship at speed c and approaches the Earth at speed c. The speed of light is independent of the relative motion of source and observer, whether the observer is on the ship or Earth-bound.

Velocities cannot add to greater than the speed of light, provided that v is less than c and u' does not exceed c. The following example illustrates that relativistic velocity addition is not as symmetric as classical velocity addition.

Example:

Comparing the Speed of Light towards and away from an Observer: Relativistic Package Delivery

Suppose the spaceship in the previous example is approaching the Earth at half the speed of light and shoots a canister at a speed of $0.750c$. (a) At what velocity will an Earth-bound observer see the canister if it is shot directly towards the Earth? (b) If it is shot directly away from the Earth? (See [link].)



Strategy

Because the canister and the spaceship are moving at relativistic speeds, we must determine the speed of the canister by an Earth-bound observer using relativistic velocity addition instead of simple velocity addition.

Solution for (a)

1. Identify the knowns. v$=0.500c$; $u\prime= 0.750c$
2. Identify the unknown. u
3. Choose the appropriate equation. u$=\frac{v+u\prime}{1+\frac{vu\prime}{c^2}}$
4. Plug the knowns into the equation.
   Equation:

$$\begin{aligned} u &= \frac{v+u\prime}{1+\frac{vu\prime}{c^2}} \\ &= \frac{0.500\text{c} + 0.750c}{1+\frac{(0.500c)(0.750c)}{c^2}} \\ &= \frac{1.250c}{1+0.375} \\ &= 0.909c \end{aligned}$$

Solution for (b)

1. Identify the knowns. $v = 0.500c$; $u\prime = -0.750c$
2. Identify the unknown. u
3. Choose the appropriate equation. $u = \frac{v+u\prime}{1+\frac{vu\prime}{c^2}}$
4. Plug the knowns into the equation.
   Equation:

$$\begin{aligned} u &= \frac{v+u\prime}{1+\frac{vu\prime}{c^2}} \\ &= \frac{0.500c+(-0.750c)}{1+\frac{(0.500c)(-0.750c)}{c^2}} \\ &= \frac{-0.250c}{1-0.375} \\ &= -0.400c \end{aligned}$$

Discussion
The minus sign indicates velocity away from the Earth (in the opposite direction from v), which means the canister is heading towards the Earth in part (a) and away in part (b), as expected. But relativistic velocities do not add as simply as they do classically. In part (a), the canister does approach the Earth faster, but not at the simple sum of $1.250c$. The total velocity is less than you would get classically. And in part (b), the canister moves away from the Earth at a velocity of $-0.400c$, which is *faster* than the $-0.250c$ you would expect classically. The velocities are not even symmetric. In part (a) the canister moves $0.409c$ faster than the ship relative to the Earth, whereas in part (b) it moves $0.900c$ slower than the ship.

Doppler Shift

Although the speed of light does not change with relative velocity, the frequencies and wavelengths of light do. First discussed for sound waves, a Doppler shift occurs in any wave when there is relative motion between source and observer.

Note:
Relativistic Doppler Effects
The observed wavelength of electromagnetic radiation is longer (called a red shift) than that emitted by the source when the source moves away

from the observer and shorter (called a blue shift) when the source moves towards the observer.

Equation:

$$=\lambda_{\text{obs}}=\lambda_s\sqrt{\frac{1+\frac{u}{c}}{1-\frac{u}{c}}}.$$

In the Doppler equation, λ_{obs} is the observed wavelength, λ_s is the source wavelength, and u is the relative velocity of the source to the observer. The velocity u is positive for motion away from an observer and negative for motion toward an observer. In terms of source frequency and observed frequency, this equation can be written

Equation:

$$f_{\text{obs}}=\mathrm{f}_s\sqrt{\frac{1-\frac{u}{c}}{1+\frac{u}{c}}}.$$

Notice that the – and + signs are different than in the wavelength equation.

Note:

Career Connection: Astronomer

If you are interested in a career that requires a knowledge of special relativity, there’s probably no better connection than astronomy. Astronomers must take into account relativistic effects when they calculate distances, times, and speeds of black holes, galaxies, quasars, and all other astronomical objects. To have a career in astronomy, you need at least an undergraduate degree in either physics or astronomy, but a Master’s or doctoral degree is often required. You also need a good background in high-level mathematics.

Example:

Calculating a Doppler Shift: Radio Waves from a Receding Galaxy

Suppose a galaxy is moving away from the Earth at a speed $0.825c$. It emits radio waves with a wavelength of 0.525 m. What wavelength would we detect on the Earth?

Strategy

Because the galaxy is moving at a relativistic speed, we must determine the Doppler shift of the radio waves using the relativistic Doppler shift instead of the classical Doppler shift.

Solution

1. Identify the knowns. u=$0.825c$; $\lambda_s = 0.525\ m$
2. Identify the unknown. λ_{obs}
3. Choose the appropriate equation. λ_{obs}=$\lambda_s\sqrt{\frac{1+\frac{u}{c}}{1-\frac{u}{c}}}$
4. Plug the knowns into the equation.

   Equation:

$$\begin{aligned}\lambda_{\text{obs}} &= \lambda_s\sqrt{\frac{1+\frac{u}{c}}{1-\frac{u}{c}}}\\ &= (0.525\text{ m})\sqrt{\frac{1+\frac{0.825\text{c}}{c}}{1-\frac{0.825\text{c}}{c}}}\\ &= 1.70\text{ m.}\end{aligned}$$

Discussion

Because the galaxy is moving away from the Earth, we expect the wavelengths of radiation it emits to be redshifted. The wavelength we calculated is 1.70 m, which is redshifted from the original wavelength of 0.525 m.

The relativistic Doppler shift is easy to observe. This equation has everyday applications ranging from Doppler-shifted radar velocity measurements of transportation to Doppler-radar storm monitoring. In astronomical observations, the relativistic Doppler shift provides velocity information such as the motion and distance of stars.

Exercise:

Check Your Understanding

Problem:

Suppose a space probe moves away from the Earth at a speed $0.350c$. It sends a radio wave message back to the Earth at a frequency of 1.50 GHz. At what frequency is the message received on the Earth?

Solution:

Answer

Equation:

$$f_{\text{obs}}{=}\mathrm{f}_s\sqrt{\frac{1-\frac{u}{c}}{1+\frac{u}{c}}}=(1.50\ \text{GHz})\sqrt{\frac{1-\frac{0.350\text{c}}{c}}{1+\frac{0.350\text{c}}{c}}}=1.04\ \text{GHz}$$

Section Summary

- With classical velocity addition, velocities add like regular numbers in one-dimensional motion: $u{=}v{+}u\prime$, where v is the velocity between two observers, u is the velocity of an object relative to one observer, and $u\prime$ is the velocity relative to the other observer.
- Velocities cannot add to be greater than the speed of light. Relativistic velocity addition describes the velocities of an object moving at a relativistic speed:
  Equation:

$$\mathrm{u}{=}\frac{v{+}u\prime}{1+\frac{vu\prime}{c^2}}$$

- An observer of electromagnetic radiation sees **relativistic Doppler effects** if the source of the radiation is moving relative to the observer. The wavelength of the radiation is longer (called a red shift) than that emitted by the source when the source moves away from the observer and shorter (called a blue shift) when the source moves toward the observer. The shifted wavelength is described by the equation

Equation:

$$\lambda_{\text{obs}}=\lambda_s\sqrt{\frac{1+\frac{u}{c}}{1-\frac{u}{c}}}$$

λ_{obs} is the observed wavelength, λ_s is the source wavelength, and u is the relative velocity of the source to the observer.

Conceptual Questions

Exercise:

Problem:

Explain the meaning of the terms “red shift” and “blue shift” as they relate to the relativistic Doppler effect.

Exercise:

Problem:

What happens to the relativistic Doppler effect when relative velocity is zero? Is this the expected result?

Exercise:

Problem:

Is the relativistic Doppler effect consistent with the classical Doppler effect in the respect that λ_{obs} is larger for motion away?

Exercise:

Problem:

All galaxies farther away than about 50×10^6 ly exhibit a red shift in their emitted light that is proportional to distance, with those farther and farther away having progressively greater red shifts. What does this imply, assuming that the only source of red shift is relative motion? (Hint: At these large distances, it is space itself that is expanding, but the effect on light is the same.)

Problems & Exercises

Exercise:

Problem:

Suppose a spaceship heading straight towards the Earth at $0.750c$ can shoot a canister at $0.500c$ relative to the ship. (a) What is the velocity of the canister relative to the Earth, if it is shot directly at the Earth? (b) If it is shot directly away from the Earth?

Solution:

(a) $0.909c$

(b) $0.400c$

Exercise:

Problem:

Repeat the previous problem with the ship heading directly away from the Earth.

Exercise:

Problem:

If a spaceship is approaching the Earth at $0.100c$ and a message capsule is sent toward it at $0.100c$ relative to the Earth, what is the speed of the capsule relative to the ship?

Solution:

$0.198c$

Exercise:

Problem:

(a) Suppose the speed of light were only 3000 m/s. A jet fighter moving toward a target on the ground at 800 m/s shoots bullets, each having a muzzle velocity of 1000 m/s. What are the bullets' velocity relative to the target? (b) If the speed of light was this small, would you observe relativistic effects in everyday life? Discuss.

Exercise:

Problem:

If a galaxy moving away from the Earth has a speed of 1000 km/s and emits 656 nm light characteristic of hydrogen (the most common element in the universe). (a) What wavelength would we observe on the Earth? (b) What type of electromagnetic radiation is this? (c) Why is the speed of the Earth in its orbit negligible here?

Solution:

a) 658 nm

b) red

c) $v/\text{c} = 9.92 \times 10^{-5}$ (negligible)

Exercise:

Problem:

A space probe speeding towards the nearest star moves at $0.250c$ and sends radio information at a broadcast frequency of 1.00 GHz. What frequency is received on the Earth?

Exercise:

Problem:

If two spaceships are heading directly towards each other at $0.800c$, at what speed must a canister be shot from the first ship to approach the other at $0.999c$ as seen by the second ship?

Solution:

$0.991c$

Exercise:

Problem:

Two planets are on a collision course, heading directly towards each other at $0.250c$. A spaceship sent from one planet approaches the second at $0.750c$ as seen by the second planet. What is the velocity of the ship relative to the first planet?

Exercise:

Problem:

When a missile is shot from one spaceship towards another, it leaves the first at $0.950c$ and approaches the other at $0.750c$. What is the relative velocity of the two ships?

Solution:

$-0.696c$

Exercise:

Problem:

What is the relative velocity of two spaceships if one fires a missile at the other at $0.750c$ and the other observes it to approach at $0.950c$?

Exercise:

Problem:

Near the center of our galaxy, hydrogen gas is moving directly away from us in its orbit about a black hole. We receive 1900 nm electromagnetic radiation and know that it was 1875 nm when emitted by the hydrogen gas. What is the speed of the gas?

Solution:

$0.01324c$

Exercise:

Problem:

A highway patrol officer uses a device that measures the speed of vehicles by bouncing radar off them and measuring the Doppler shift. The outgoing radar has a frequency of 100 GHz and the returning echo has a frequency 15.0 kHz higher. What is the velocity of the vehicle? Note that there are two Doppler shifts in echoes. Be certain not to round off until the end of the problem, because the effect is small.

Exercise:

Problem:

Prove that for any relative velocity v between two observers, a beam of light sent from one to the other will approach at speed c (provided that v is less than c, of course).

Solution:

$u\prime = c$, so

$$\begin{aligned} u &= \frac{\text{v+u}\prime}{1+(\text{vu}\prime/c^2)} = \frac{\text{v+c}}{1+(\text{vc}/c^2)} = \frac{\text{v+c}}{1+(v/c)} \\ &= \frac{c(\text{v+c})}{\text{c+v}} = c \end{aligned}$$

Exercise:

Problem:

Show that for any relative velocity v between two observers, a beam of light projected by one directly away from the other will move away at the speed of light (provided that v is less than c, of course).

Exercise:

Problem:

(a) All but the closest galaxies are receding from our own Milky Way Galaxy. If a galaxy 12.0×10^9 ly ly away is receding from us at 0. $0.900c$, at what velocity relative to us must we send an exploratory probe to approach the other galaxy at $0.990c$, as measured from that galaxy? (b) How long will it take the probe to reach the other galaxy as measured from the Earth? You may assume that the velocity of the other galaxy remains constant. (c) How long will it then take for a radio signal to be beamed back? (All of this is possible in principle, but not practical.)

Solution:

a) $0.99947c$

b) 1.2064×10^{11} y

c) 1.2058×10^{11} y (all to sufficient digits to show effects)

Glossary

classical velocity addition
: the method of adding velocities when $v<<c$; velocities add like regular numbers in one-dimensional motion: $u = v+u\prime$, where v is the

velocity between two observers, u is the velocity of an object relative to one observer, and $u\prime$ is the velocity relative to the other observer

relativistic velocity addition

the method of adding velocities of an object moving at a relativistic speed: $\text{u}=\frac{v+u\prime}{1+\frac{vu\prime}{c^2}}$, where v is the relative velocity between two observers, u is the velocity of an object relative to one observer, and $u\prime$ is the velocity relative to the other observer

relativistic Doppler effects

a change in wavelength of radiation that is moving relative to the observer; the wavelength of the radiation is longer (called a red shift) than that emitted by the source when the source moves away from the observer and shorter (called a blue shift) when the source moves toward the observer; the shifted wavelength is described by the equation

Equation:

$$\lambda_{\text{obs}}=\lambda_s\sqrt{\frac{1+\frac{u}{c}}{1-\frac{u}{c}}}$$

where λ_{obs} is the observed wavelength, λ_s is the source wavelength, and u is the velocity of the source to the observer

Relativistic Momentum

- Calculate relativistic momentum.
- Explain why the only mass it makes sense to talk about is rest mass.

Momentum is an important concept for these football players from the University of California at Berkeley and the University of California at Davis. Players with more mass often have a larger impact because their momentum is larger. For objects moving at relativistic speeds, the effect is even greater. (credit: John Martinez Pavliga)

In classical physics, momentum is a simple product of mass and velocity. However, we saw in the last section that when special relativity is taken into account, massive objects have a speed limit. What effect do you think mass and velocity have on the momentum of objects moving at relativistic speeds?

Momentum is one of the most important concepts in physics. The broadest form of Newton’s second law is stated in terms of momentum. Momentum is conserved whenever the net external force on a system is zero. This makes momentum conservation a fundamental tool for analyzing collisions. All of Work, Energy, and Energy Resources is devoted to momentum, and momentum has been important for many other topics as well, particularly where collisions were involved. We will see that momentum has the same importance in modern physics. Relativistic momentum is conserved, and much of what we know about subatomic structure comes from the analysis of collisions of accelerator-produced relativistic particles.

The first postulate of relativity states that the laws of physics are the same in all inertial frames. Does the law of conservation of momentum survive this requirement at high velocities? The answer is yes, provided that the momentum is defined as follows.

Note:

Relativistic Momentum

Relativistic momentum p is classical momentum multiplied by the relativistic factor γ.

Equation:

$$p = \gamma mu,$$

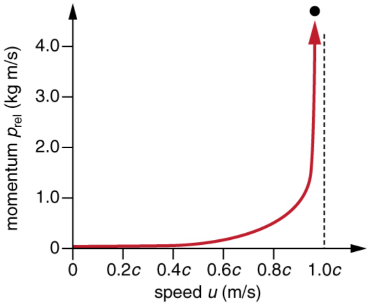
where m is the **rest mass** of the object, u is its velocity relative to an observer, and the relativistic factor

Equation:

$$\gamma = \frac{1}{\sqrt{1 - \frac{u^2}{c^2}}}.$$

Note that we use u for velocity here to distinguish it from relative velocity v between observers. Only one observer is being considered here. With p defined in this way, total momentum p_{tot} is conserved whenever the net external force is zero, just as in classical physics. Again we see that the relativistic quantity becomes virtually the same as the classical at low velocities. That is, relativistic momentum γmu becomes the classical mu at low velocities, because γ is very nearly equal to 1 at low velocities.

Relativistic momentum has the same intuitive feel as classical momentum. It is greatest for large masses moving at high velocities, but, because of the factor γ, relativistic momentum approaches infinity as u approaches c. (See [link].) This is another indication that an object with mass cannot reach the speed of light. If it did, its momentum would become infinite, an unreasonable value.



Relativistic momentum approaches infinity as the velocity of an object approaches the speed of light.

Note:

Misconception Alert: Relativistic Mass and Momentum

The relativistically correct definition of momentum as $p = \gamma mu$ is sometimes taken to imply that mass varies with velocity: $m_{\text{var}} = \gamma m$, particularly in older textbooks. However, note that m is the mass of the object as measured by a person at rest relative to the object. Thus, m is defined to be the rest mass, which could be measured at rest, perhaps using gravity. When a mass is moving relative to an observer, the only way that its mass can be determined is through collisions or other means in which momentum is involved. Since the mass of a moving object cannot be determined independently of momentum, the only meaningful mass is rest mass. Thus, when we use the term mass, assume it to be identical to rest mass.

Relativistic momentum is defined in such a way that the conservation of momentum will hold in all inertial frames. Whenever the net external force on a system is zero, relativistic momentum is conserved, just as is the case for classical momentum. This has been verified in numerous experiments.

In Relativistic Energy, the relationship of relativistic momentum to energy is explored. That subject will produce our first inkling that objects without mass may also have momentum.

Exercise:
Check Your Understanding

Problem:

What is the momentum of an electron traveling at a speed $0.985c$? The rest mass of the electron is $9.11\times10^{-31}\text{ kg}$.

Solution:
Answer
Equation:

$$p=\gamma\text{mu}=\frac{\text{mu}}{\sqrt{1-\frac{u^2}{c^2}}}==\frac{(9.11\times10^{-31}\text{ kg})(0.985)(3.00\times10^{8}\text{ m/s})}{\sqrt{1-\frac{(0.985c)^2}{c^2}}}=1.56\times10^{-21}\text{ kg}\cdot\text{m/s}$$

Section Summary

- The law of conservation of momentum is valid whenever the net external force is zero and for relativistic momentum. Relativistic momentum p is classical momentum multiplied by the relativistic factor γ.
- $p=\gamma\text{mu}$, where m is the rest mass of the object, u is its velocity relative to an observer, and the relativistic factor $\gamma=\frac{1}{\sqrt{1-\frac{u^2}{c^2}}}$.
- At low velocities, relativistic momentum is equivalent to classical momentum.
- Relativistic momentum approaches infinity as u approaches c. This implies that an object with mass cannot reach the speed of light.
- Relativistic momentum is conserved, just as classical momentum is conserved.

Conceptual Questions

Exercise:

Problem: How does modern relativity modify the law of conservation of momentum?

Exercise:

Problem:

Is it possible for an external force to be acting on a system and relativistic momentum to be conserved? Explain.

Problem Exercises

Exercise:

Problem:

Find the momentum of a helium nucleus having a mass of 6.68×10^{-27} kg that is moving at $0.200c$.

Solution:

$4.09 \times 10^{-19} \text{ kg} \cdot \text{m/s}$

Exercise:

Problem: What is the momentum of an electron traveling at $0.980c$?

Exercise:

Problem:

(a) Find the momentum of a 1.00×10^{9} kg asteroid heading towards the Earth at 30.0 km/s. (b) Find the ratio of this momentum to the classical momentum. (Hint: Use the approximation that $\gamma = 1 + (1/2)v^2/c^2$ at low velocities.)

Solution:

(a) $3.000000015 \times 10^{13} \text{ kg} \cdot \text{m/s}$.

(b) Ratio of relativistic to classical momenta equals 1.000000005 (extra digits to show small effects)

Exercise:

Problem:

(a) What is the momentum of a 2000 kg satellite orbiting at 4.00 km/s? (b) Find the ratio of this momentum to the classical momentum. (Hint: Use the approximation that $\gamma = 1 + (1/2)v^2/c^2$ at low velocities.)

Exercise:

Problem:

What is the velocity of an electron that has a momentum of 3.04×10^{-21} kg$\cdot$m/s? Note that you must calculate the velocity to at least four digits to see the difference from c.

Solution:

$2.9957 \times 10^{8}\ \mathrm{m/s}$

Exercise:

Problem: Find the velocity of a proton that has a momentum of $4.48 \times –10^{-19}\ \mathrm{kg{\cdot}m/s}$.

Exercise:

Problem:

(a) Calculate the speed of a 1.00-μg particle of dust that has the same momentum as a proton moving at $0.999c$. (b) What does the small speed tell us about the mass of a proton compared to even a tiny amount of macroscopic matter?

Solution:

(a) $1.121 \times 10^{-8}\ \mathrm{m/s}$

(b) The small speed tells us that the mass of a proton is substantially smaller than that of even a tiny amount of macroscopic matter!

Exercise:

Problem:

(a) Calculate γ for a proton that has a momentum of $1.00\ \mathrm{kg{\cdot}m/s}$. (b) What is its speed? Such protons form a rare component of cosmic radiation with uncertain origins.

Glossary

relativistic momentum

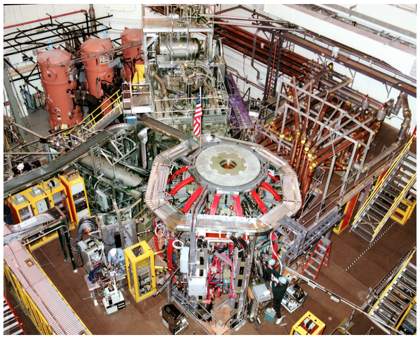
p, the momentum of an object moving at relativistic velocity; $p = \mathrm{\gamma mu}$, where m is the rest mass of the object, u is its velocity relative to an observer, and the relativistic factor $\gamma = \frac{1}{\sqrt{1-\frac{u^2}{c^2}}}$

rest mass

the mass of an object as measured by a person at rest relative to the object

Relativistic Energy

- Compute total energy of a relativistic object.
- Compute the kinetic energy of a relativistic object.
- Describe rest energy, and explain how it can be converted to other forms.
- Explain why massive particles cannot reach C.



The National Spherical Torus Experiment (NSTX) has a fusion reactor in which hydrogen isotopes undergo fusion to produce helium. In this process, a relatively small mass of fuel is converted into a large amount of energy. (credit: Princeton Plasma Physics Laboratory)

A tokamak is a form of experimental fusion reactor, which can change mass to energy. Accomplishing this requires an understanding of relativistic energy. Nuclear reactors are proof of the conservation of relativistic energy.

Conservation of energy is one of the most important laws in physics. Not only does energy have many important forms, but each form can be converted to any other. We know that classically the total amount of energy in a system remains constant. Relativistically, energy is still conserved, provided its definition is altered to include the possibility of mass changing to energy, as in the reactions that occur within a nuclear reactor. Relativistic energy is intentionally defined so that it will be conserved in all inertial frames, just as is the case for relativistic momentum. As a consequence, we learn that several fundamental quantities are related in ways not known in classical physics. All of these relationships are verified by experiment and have fundamental consequences. The altered definition of energy contains some of the most fundamental and spectacular new insights into nature found in recent history.

Total Energy and Rest Energy

The first postulate of relativity states that the laws of physics are the same in all inertial frames. Einstein showed that the law of conservation of energy is valid relativistically, if we define energy to include a relativistic factor.

Note:
Total Energy
Total energy E is defined to be
Equation:

$$E=\gamma mc^2,$$

where m is mass, c is the speed of light, $\gamma=\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$, and v is the velocity of the mass relative to an observer. There are many aspects of the total energy E that we will discuss—among them are how kinetic and potential energies are included in E, and how E is related to relativistic momentum. But first, note that at rest, total energy is not zero. Rather, when $v=0$, we have $\gamma=1$, and an object has rest energy.

Note:
Rest Energy
Rest energy is
Equation:

$$E_0=mc^2.$$

This is the correct form of Einstein’s most famous equation, which for the first time showed that energy is related to the mass of an object at rest. For example, if energy is stored in the object, its rest mass increases. This also implies that mass can be destroyed to release energy. The implications of these first two equations regarding relativistic energy are so broad that they were not completely recognized for some years after Einstein published them in 1907, nor was the experimental proof that they are correct widely recognized at first. Einstein, it should be noted, did understand and describe the meanings and implications of his theory.

Example:
Calculating Rest Energy: Rest Energy is Very Large

Calculate the rest energy of a 1.00-g mass.

Strategy

One gram is a small mass—less than half the mass of a penny. We can multiply this mass, in SI units, by the speed of light squared to find the equivalent rest energy.

Solution

1. Identify the knowns. $m = 1.00 \times 10^{-3}\ \text{kg}$; $c = 3.00 \times 10^{8}\ \text{m/s}$
2. Identify the unknown. E_0
3. Choose the appropriate equation. $E_0 = mc^2$
4. Plug the knowns into the equation.
   Equation:

$$\begin{aligned} E_0 &= mc^2 = (1.00 \times 10^{-3}\ \text{kg})(3.00 \times 10^{8}\ \text{m/s})^2 \\ &= 9.00 \times 10^{13}\ \text{kg} \cdot \text{m}^2/\text{s}^2 \end{aligned}$$

5. Convert units.

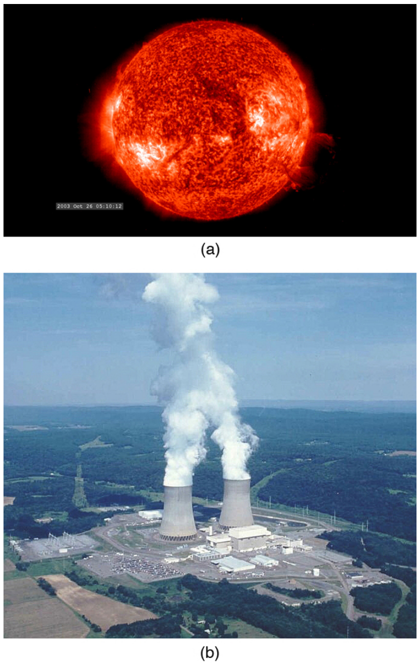
   Noting that $1\ \text{kg} \cdot \text{m}^2/\text{s}^2 = 1\ \text{J}$, we see the rest mass energy is
   Equation:

$$E_0 = 9.00 \times 10^{13}\ \text{J}.$$

Discussion

This is an enormous amount of energy for a 1.00-g mass. We do not notice this energy, because it is generally not available. Rest energy is large because the speed of light c is a large number and c^2 is a very large number, so that mc^2 is huge for any macroscopic mass. The 9.00×10^{13} J rest mass energy for 1.00 g is about twice the energy released by the Hiroshima atomic bomb and about 10,000 times the kinetic energy of a large aircraft carrier. If a way can be found to convert rest mass energy into some other form (and all forms of energy can be converted into one another), then huge amounts of energy can be obtained from the destruction of mass.

Today, the practical applications of *the conversion of mass into another form of energy*, such as in nuclear weapons and nuclear power plants, are well known. But examples also existed when Einstein first proposed the correct form of relativistic energy, and he did describe some of them. Nuclear radiation had been discovered in the previous decade, and it had been a mystery as to where its energy originated. The explanation was that, in certain nuclear processes, a small amount of mass is destroyed and energy is released and carried by nuclear radiation. But the amount of mass destroyed is so small that it is difficult to detect that any is missing. Although Einstein proposed this as the source of energy in the radioactive salts then being studied, it was many years before there was broad recognition that mass could be and, in fact, commonly is converted to energy. (See [link].)



(a)

(b)

The Sun (a) and the Susquehanna Steam Electric Station (b) both convert mass into energy—the Sun via nuclear fusion, the electric station via nuclear fission. (credits: (a) NASA/Goddard Space Flight Center, Scientific Visualization Studio; (b) U.S. government)

Because of the relationship of rest energy to mass, we now consider mass to be a form of energy rather than something separate. There had not even been a hint of this prior to Einstein’s work. Such conversion is now known to be the source of the Sun’s energy, the energy of nuclear decay, and even the source of energy keeping Earth’s interior hot.

Stored Energy and Potential Energy

What happens to energy stored in an object at rest, such as the energy put into a battery by charging it, or the energy stored in a toy gun’s compressed spring? The energy input becomes part of the total energy of the object and, thus, increases its rest mass. All stored and potential energy becomes mass in a system. Why is it we don’t ordinarily notice this? In fact, conservation of mass (meaning total mass is constant) was one of the great laws verified by 19th-century science. Why was it not noticed to be incorrect? The following example helps answer these questions.

Example:
Calculating Rest Mass: A Small Mass Increase due to Energy Input

A car battery is rated to be able to move 600 ampere-hours $(\mathrm{A{\cdot}h})$ of charge at 12.0 V. (a) Calculate the increase in rest mass of such a battery when it is taken from being fully depleted to being fully charged. (b) What percent increase is this, given the battery’s mass is 20.0 kg?

Strategy

In part (a), we first must find the energy stored in the battery, which equals what the battery can supply in the form of electrical potential energy. Since $\mathrm{PE_{elec}} = \mathrm{qV}$, we have to calculate the charge q in $600\ \mathrm{A{\cdot}h}$, which is the product of the current I and the time t. We then multiply the result by 12.0 V. We can then calculate the battery’s increase in mass using $\Delta E = \mathrm{PE_{elec}} = (\Delta m)c^2$. Part (b) is a simple ratio converted to a percentage.

Solution for (a)

1. Identify the knowns. $I \cdot t = 600\ \mathrm{A \cdot h}$; $V = 12.0\ \mathrm{V}$; $c = 3.00 \times 10^8\ \mathrm{m/s}$
2. Identify the unknown. Δm
3. Choose the appropriate equation. $\mathrm{PE_{elec}} = (\Delta m)c^2$
4. Rearrange the equation to solve for the unknown. $\Delta m = \frac{\mathrm{PE_{elec}}}{c^2}$
5. Plug the knowns into the equation.
   Equation:

$$\begin{aligned}\Delta m &= \frac{\mathrm{PE_{elec}}}{c^2}\\ &= \frac{\mathrm{qV}}{c^2}\\ &= \frac{(It)V}{c^2}\\ &= \frac{(600\ \mathrm{A{\cdot}h})(12.0\ \mathrm{V})}{(3.00{\times}10^8)^2}.\end{aligned}$$

   Write amperes A as coulombs per second (C/s), and convert hours to seconds.
   Equation:

$$\begin{aligned}\Delta m &= \frac{(600\text{ C/s}\cdot\text{h}\left(\frac{3600\text{ s}}{1\text{ h}}\right)(12.0\text{ J/C})}{(3.00\times10^{8}\text{ m/s})^{2}} \\ &= \frac{(2.16\times10^{6}\text{ C})(12.0\text{ J/C})}{(3.00\times10^{8}\text{ m/s})^{2}}\end{aligned}$$

Using the conversion $1\text{ kg}\cdot\text{m}^2/\text{s}^2 = 1\text{ J}$, we can write the mass as

$\Delta m = 2.88\times10^{-10}\text{ kg}$.

Solution for (b)

1. Identify the knowns. $\Delta m = 2.88\times10^{-10}\text{ kg}$; $m = 20.0\text{ kg}$
2. Identify the unknown. % change
3. Choose the appropriate equation. $\%\text{ increase} = \frac{\Delta m}{m}\times100\%$
4. Plug the knowns into the equation.
   Equation:

$$\begin{aligned}\%\text{ increase} &= \frac{\Delta m}{m}\times100\% \\ &= \frac{2.88\times10^{-10}\text{ kg}}{20.0\text{ kg}}\times100\% \\ &= 1.44\times10^{-9}\%.\end{aligned}$$

Discussion
Both the actual increase in mass and the percent increase are very small, since energy is divided by c^2, a very large number. We would have to be able to measure the mass of the battery to a precision of a billionth of a percent, or 1 part in 10^{11}, to notice this increase. It is no wonder that the mass variation is not readily observed. In fact, this change in mass is so small that we may question how you could verify it is real. The answer is found in nuclear processes in which the percentage of mass destroyed is large enough to be measured. The mass of the fuel of a nuclear reactor, for example, is measurably smaller when its energy has been used. In that case, stored energy has been released (converted mostly to heat and electricity) and the rest mass has decreased. This is also the case when you use the energy stored in a battery, except that the stored energy is much greater in nuclear processes, making the change in mass measurable in practice as well as in theory.

Kinetic Energy and the Ultimate Speed Limit

Kinetic energy is energy of motion. Classically, kinetic energy has the familiar expression $\frac{1}{2}mv^2$. The relativistic expression for kinetic energy is obtained from the work-energy theorem. This theorem states that the net work on a system goes into kinetic energy. If our system starts from rest, then the work-energy theorem is
Equation:

$$W_{\text{net}} = \text{KE}.$$

Relativistically, at rest we have rest energy $E_0 = mc^2$. The work increases this to the total energy $E = \gamma mc^2$. Thus,

Equation:

$$W_{\text{net}} = E - E_0 = \gamma mc^2 - mc^2 = (\gamma - 1)\,mc^2.$$

Relativistically, we have $W_{\text{net}} = \text{KE}_{\text{rel}}$.

Note:
Relativistic Kinetic Energy
Relativistic kinetic energy is
Equation:

$$\text{KE}_{\text{rel}} = (\gamma - 1)\,mc^2.$$

When motionless, we have $v = 0$ and

Equation:

$$\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} = 1,$$

so that $\text{KE}_{\text{rel}} = 0$ at rest, as expected. But the expression for relativistic kinetic energy (such as total energy and rest energy) does not look much like the classical $\frac{1}{2}mv^2$. To show that the classical expression for kinetic energy is obtained at low velocities, we note that the binomial expansion for γ at low velocities gives

Equation:

$$\gamma = 1 + \frac{1}{2}\frac{v^2}{c^2}.$$

A binomial expansion is a way of expressing an algebraic quantity as a sum of an infinite series of terms. In some cases, as in the limit of small velocity here, most terms are very small. Thus the expression derived for γ here is not exact, but it is a very accurate approximation. Thus, at low velocities,

Equation:

$$\gamma - 1 = \frac{1}{2}\frac{v^2}{c^2}.$$

Entering this into the expression for relativistic kinetic energy gives
Equation:

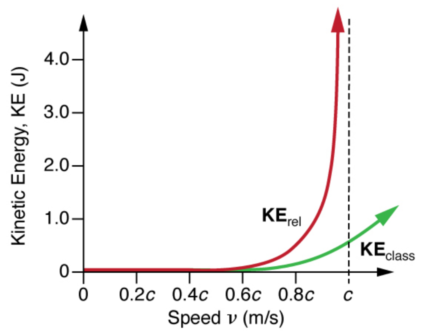
$$\mathrm{KE_{rel}} = \left[\frac{1}{2}\frac{v^2}{c^2}\right]mc^2 = \frac{1}{2}mv^2 = \mathrm{KE_{class}}.$$

So, in fact, relativistic kinetic energy does become the same as classical kinetic energy when $v<<c$.

It is even more interesting to investigate what happens to kinetic energy when the velocity of an object approaches the speed of light. We know that γ becomes infinite as v approaches c, so that $\mathrm{KE_{rel}}$ also becomes infinite as the velocity approaches the speed of light. (See [link].) An infinite amount of work (and, hence, an infinite amount of energy input) is required to accelerate a mass to the speed of light.

Note:
The Speed of Light
No object with mass can attain the speed of light.

So the speed of light is the ultimate speed limit for any particle having mass. All of this is consistent with the fact that velocities less than c always add to less than c. Both the relativistic form for kinetic energy and the ultimate speed limit being c have been confirmed in detail in numerous experiments. No matter how much energy is put into accelerating a mass, its velocity can only approach—not reach—the speed of light.



This graph of $\mathrm{KE_{rel}}$

versus velocity shows how kinetic energy approaches infinity as velocity approaches the speed of light. It is thus not possible for an object having mass to reach the speed of light. Also shown is KE_{class}, the classical kinetic energy, which is similar to relativistic kinetic energy at low velocities. Note that much more energy is required to reach high velocities than predicted classically.

Example:
Comparing Kinetic Energy: Relativistic Energy Versus Classical Kinetic Energy

An electron has a velocity $v = 0.990c$. (a) Calculate the kinetic energy in MeV of the electron. (b) Compare this with the classical value for kinetic energy at this velocity. (The mass of an electron is 9.11×10^{-31} kg.)

Strategy

The expression for relativistic kinetic energy is always correct, but for (a) it must be used since the velocity is highly relativistic (close to c). First, we will calculate the relativistic factor γ, and then use it to determine the relativistic kinetic energy. For (b), we will calculate the classical kinetic energy (which would be close to the relativistic value if v were less than a few percent of c) and see that it is not the same.

Solution for (a)

1. Identify the knowns. $v = 0.990c$; $m = 9.11 \times 10^{-31}$ kg
2. Identify the unknown. KE_{rel}
3. Choose the appropriate equation. $\text{KE}_{\text{rel}} = (\gamma - 1)\, mc^2$
4. Plug the knowns into the equation.

   First calculate γ. We will carry extra digits because this is an intermediate calculation.
   Equation:

$$\begin{aligned}\gamma &= \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}\\ &= \frac{1}{\sqrt{1-\frac{(0.990\text{c})^2}{c^2}}}\\ &= \frac{1}{\sqrt{1-(0.990)^2}}\\ &= 7.0888\end{aligned}$$

Next, we use this value to calculate the kinetic energy.
Equation:

$$\begin{aligned}\text{KE}_{\text{rel}} &= (\gamma - 1)\,mc^2\\ &= (7.0888 - 1)(9.11 \times 10^{-31}\text{ kg})(3.00 \times 10^{8}\text{ m/s})^2\\ &= 4.99 \times 10^{-13}\text{ J}\end{aligned}$$

5. Convert units.
Equation:

$$\begin{aligned}\text{KE}_{\text{rel}} &= (4.99 \times 10^{-13}\text{ J})\left(\frac{1\text{ MeV}}{1.60\times10^{-13}\text{ J}}\right)\\ &= 3.12\text{ MeV}\end{aligned}$$

Solution for (b)

1. List the knowns. $v = 0.990c$; $m = 9.11 \times 10^{-31}$ kg
2. List the unknown. KE_{class}
3. Choose the appropriate equation. $\text{KE}_{\text{class}} = \frac{1}{2}mv^2$
4. Plug the knowns into the equation.
Equation:

$$\begin{aligned}\text{KE}_{\text{class}} &= \frac{1}{2}mv^2\\ &= \frac{1}{2}(9.00 \times 10^{-31}\text{ kg})(0.990)^2(3.00 \times 10^{8}\text{ m/s})^2\\ &= 4.02 \times 10^{-14}\text{ J}\end{aligned}$$

5. Convert units.
Equation:

$$\begin{aligned}\text{KE}_{\text{class}} &= 4.02 \times 10^{-14}\text{ J}\left(\frac{1\text{ MeV}}{1.60\times10^{-13}\text{ J}}\right)\\ &= 0.251\text{ MeV}\end{aligned}$$

Discussion

As might be expected, since the velocity is 99.0% of the speed of light, the classical kinetic energy is significantly off from the correct relativistic value. Note also that the classical value is much smaller than the relativistic value. In fact, $\text{KE}_{\text{rel}}/\text{KE}_{\text{class}} = 12.4$ here. This is some indication of how difficult it is to get a mass moving close to the speed of light. Much more energy is required than predicted classically. Some people interpret this extra energy as going into increasing the mass of the system, but, as discussed in Relativistic Momentum, this cannot be verified unambiguously. What is certain is that ever-increasing amounts of energy are needed to get the velocity of a mass a little closer to that of light. An energy of 3 MeV is a very small amount for an electron, and it can be achieved with present-day particle accelerators. SLAC, for example, can accelerate electrons to over $50 \times 10^9 \text{ eV} = 50{,}000 \text{ MeV}$.

Is there any point in getting v a little closer to c than 99.0% or 99.9%? The answer is yes. We learn a great deal by doing this. The energy that goes into a high-velocity mass can be converted to any other form, including into entirely new masses. (See [link].) Most of what we know about the substructure of matter and the collection of exotic short-lived particles in nature has been learned this way. Particles are accelerated to extremely relativistic energies and made to collide with other particles, producing totally new species of particles. Patterns in the characteristics of these previously unknown particles hint at a basic substructure for all matter. These particles and some of their characteristics will be covered in Particle Physics.

The Fermi National Accelerator Laboratory, near Batavia, Illinois, was a subatomic particle collider that accelerated protons and antiprotons to attain energies up to 1 Tev (a trillion electronvolts). The circular ponds near the rings were built to dissipate waste heat. This accelerator was shut down in September 2011. (credit: Fermilab, Reidar Hahn)

Relativistic Energy and Momentum

We know classically that kinetic energy and momentum are related to each other, since
Equation:

$$\mathrm{KE}_{\text{class}} = \frac{p^2}{2m} = \frac{(mv)^2}{2m} = \frac{1}{2}mv^2.$$

Relativistically, we can obtain a relationship between energy and momentum by algebraically manipulating their definitions. This produces
Equation:

$$E^2 = (\mathrm{pc})^2 + (\mathrm{mc}^2)^2,$$

where E is the relativistic total energy and p is the relativistic momentum. This relationship between relativistic energy and relativistic momentum is more complicated than the classical, but we can gain some interesting new insights by examining it. First, total energy is related to momentum and rest mass. At rest, momentum is zero, and the equation gives the total energy to be the rest energy mc^2 (so this equation is consistent with the discussion of rest energy above). However, as the mass is accelerated, its momentum p increases, thus increasing the total energy. At sufficiently high velocities, the rest energy term $(mc^2)^2$ becomes negligible compared with the momentum term $(pc)^2$; thus, $E = pc$ at extremely relativistic velocities.

If we consider momentum p to be distinct from mass, we can determine the implications of the equation $E^2 = (\mathrm{pc})^2 + (\mathrm{mc}^2)^2$, for a particle that has no mass. If we take m to be zero in this equation, then $E = \mathrm{pc}$, or $p = E/c$. Massless particles have this momentum. There are several massless particles found in nature, including photons (these are quanta of electromagnetic radiation). Another implication is that a massless particle must travel at speed c and only at speed c. While it is beyond the scope of this text to examine the relationship in the equation $E^2 = (\mathrm{pc})^2 + (\mathrm{mc}^2)^2$, in detail, we can see that the relationship has important implications in special relativity.

Note:
Problem-Solving Strategies for Relativity

Examine the situation to	. Relativistic	$\gamma = \frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$	, the quantitative	γ is very close to 1, then relativistic effects are small and differ very

determine that it is necessary to use relativity	effects are related to relativistic factor. If little from the usually easier classical calculations.
Identify exactly what needs to be determined in the problem (identify the unknowns).	
Make a list of what is given or can be inferred from the problem as stated (identify the knowns).	Look in particular for information on relative velocity v.
Make certain you understand the conceptual aspects of the problem before making any calculations.	Decide, for example, which observer sees time dilated or length contracted before plugging into equations. If you have thought about who sees what, who is moving with the event being observed, who sees proper time, and so on, you will find it much easier to determine if your calculation is reasonable.
Determine the primary type of calculation to be done to find the unknowns identified above.	You will find the section summary helpful in determining whether a length contraction, relativistic kinetic energy, or some other concept is involved.
Do not round off during the calculation.	As noted in the text, you must often perform your calculations to many digits to see the desired effect. You may round off at the very end of the problem, but do not use a rounded number in a subsequent calculation.
Check the answer to see if it is reasonable: Does it make sense?	This may be more difficult for relativity, since we do not encounter it directly. But you can look for velocities greater than c or relativistic effects that are in the wrong direction (such as a time contraction where a dilation was expected).

Exercise:
Check Your Understanding

Problem:

A photon decays into an electron-positron pair. What is the kinetic energy of the electron if its speed is $0.992c$?

Solution:
Answer
Equation:

$$\begin{aligned}\text{KE}_{\text{rel}} &= (\gamma-1)\,mc^2 = \left(\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}-1\right)mc^2\\ &= \left(\frac{1}{\sqrt{1-\frac{(0.992c)^2}{c^2}}}-1\right)(9.11\times10^{-31}\text{ kg})(3.00\times10^{8}\text{ m/s})^2 = 5.67\times10^{-13}\text{ J}\end{aligned}$$

Section Summary

- Relativistic energy is conserved as long as we define it to include the possibility of mass changing to energy.
- Total Energy is defined as: $E = \gamma mc^2$, where $\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$.
- Rest energy is $E_0 = mc^2$, meaning that mass is a form of energy. If energy is stored in an object, its mass increases. Mass can be destroyed to release energy.
- We do not ordinarily notice the increase or decrease in mass of an object because the change in mass is so small for a large increase in energy.
- The relativistic work-energy theorem is $W_{\text{net}} = E - E_0 = \gamma mc^2 - mc^2 = (\gamma - 1) mc^2$.
- Relativistically, $W_{\text{net}} = \text{KE}_{\text{rel}}$, where KE_{rel} is the relativistic kinetic energy.
- Relativistic kinetic energy is $\text{KE}_{\text{rel}} = (\gamma - 1) mc^2$, where $\gamma = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$. At low velocities, relativistic kinetic energy reduces to classical kinetic energy.
- **No object with mass can attain the speed of light** because an infinite amount of work and an infinite amount of energy input is required to accelerate a mass to the speed of light.
- The equation $E^2 = (pc)^2 + (mc^2)^2$ relates the relativistic total energy E and the relativistic momentum p. At extremely high velocities, the rest energy mc^2 becomes negligible, and $E = pc$.

Conceptual Questions

Exercise:

Problem:

How are the classical laws of conservation of energy and conservation of mass modified by modern relativity?

Exercise:

Problem:

What happens to the mass of water in a pot when it cools, assuming no molecules escape or are added? Is this observable in practice? Explain.

Exercise:

Problem:

Consider a thought experiment. You place an expanded balloon of air on weighing scales outside in the early morning. The balloon stays on the scales and you are able to measure changes in its mass. Does the mass of the balloon change as the day progresses? Discuss the difficulties in carrying out this experiment.

Exercise:

Problem:

The mass of the fuel in a nuclear reactor decreases by an observable amount as it puts out energy. Is the same true for the coal and oxygen combined in a conventional power plant? If so, is this observable in practice for the coal and oxygen? Explain.

Exercise:

Problem:

We know that the velocity of an object with mass has an upper limit of c. Is there an upper limit on its momentum? Its energy? Explain.

Exercise:

Problem: Given the fact that light travels at c, can it have mass? Explain.

Exercise:

Problem:

If you use an Earth-based telescope to project a laser beam onto the Moon, you can move the spot across the Moon's surface at a velocity greater than the speed of light. Does this violate modern relativity? (Note that light is being sent from the Earth to the Moon, not across the surface of the Moon.)

Problems & Exercises

Exercise:

Problem:

What is the rest energy of an electron, given its mass is 9.11×10^{-31} kg? Give your answer in joules and MeV.

Solution:

8.20×10^{-14} J

0.512 MeV

Exercise:

Problem:

Find the rest energy in joules and MeV of a proton, given its mass is 1.67×10^{-27} kg.

Exercise:

Problem:

If the rest energies of a proton and a neutron (the two constituents of nuclei) are 938.3 and 939.6 MeV respectively, what is the difference in their masses in kilograms?

Solution:

$2.3 \times 10^{-30}\text{ kg}$

Exercise:

Problem:

The Big Bang that began the universe is estimated to have released 10^{68} J of energy. How many stars could half this energy create, assuming the average star's mass is $4.00 \times 10^{30}\text{ kg}$?

Exercise:

Problem:

A supernova explosion of a 2.00×10^{31} kg star produces 1.00×10^{44} J of energy. (a) How many kilograms of mass are converted to energy in the explosion? (b) What is the ratio $\Delta m / m$ of mass destroyed to the original mass of the star?

Solution:

(a) $1.11 \times 10^{27}\text{ kg}$

(b) 5.56×10^{-5}

Exercise:

Problem:

(a) Using data from [link], calculate the mass converted to energy by the fission of 1.00 kg of uranium. (b) What is the ratio of mass destroyed to the original mass, $\Delta m / m$?

Exercise:

Problem:

(a) Using data from [link], calculate the amount of mass converted to energy by the fusion of 1.00 kg of hydrogen. (b) What is the ratio of mass destroyed to the original mass, $\Delta m / m$? (c) How does this compare with $\Delta m / m$ for the fission of 1.00 kg of uranium?

Solution:

$7.1 \times 10^{-3}\text{ kg}$

7.1×10^{-3}

The ratio is greater for hydrogen.

Exercise:

Problem:

There is approximately 10^{34} J of energy available from fusion of hydrogen in the world’s oceans. (a) If 10^{33} J of this energy were utilized, what would be the decrease in mass of the oceans? Assume that 0.08% of the mass of a water molecule is converted to energy during the fusion of hydrogen. (b) How great a volume of water does this correspond to? (c) Comment on whether this is a significant fraction of the total mass of the oceans.

Exercise:

Problem:

A muon has a rest mass energy of 105.7 MeV, and it decays into an electron and a massless particle. (a) If all the lost mass is converted into the electron’s kinetic energy, find γ for the electron. (b) What is the electron’s velocity?

Solution:

208

$0.999988c$

Exercise:

Problem:

A π-meson is a particle that decays into a muon and a massless particle. The π-meson has a rest mass energy of 139.6 MeV, and the muon has a rest mass energy of 105.7 MeV. Suppose the π-meson is at rest and all of the missing mass goes into the muon’s kinetic energy. How fast will the muon move?

Exercise:

Problem:

(a) Calculate the relativistic kinetic energy of a 1000-kg car moving at 30.0 m/s if the speed of light were only 45.0 m/s. (b) Find the ratio of the relativistic kinetic energy to classical.

Solution:

6.92×10^{5} J

1.54

Exercise:

Problem:

Alpha decay is nuclear decay in which a helium nucleus is emitted. If the helium nucleus has a mass of 6.80×10^{-27} kg and is given 5.00 MeV of kinetic energy, what is its velocity?

Exercise:

Problem:

(a) Beta decay is nuclear decay in which an electron is emitted. If the electron is given 0.750 MeV of kinetic energy, what is its velocity? (b) Comment on how the high velocity is consistent with the kinetic energy as it compares to the rest mass energy of the electron.

Solution:

(a) $0.914c$

(b) The rest mass energy of an electron is 0.511 MeV, so the kinetic energy is approximately 150% of the rest mass energy. The electron should be traveling close to the speed of light.

Exercise:

Problem:

A positron is an antimatter version of the electron, having exactly the same mass. When a positron and an electron meet, they annihilate, converting all of their mass into energy. (a) Find the energy released, assuming negligible kinetic energy before the annihilation. (b) If this energy is given to a proton in the form of kinetic energy, what is its velocity? (c) If this energy is given to another electron in the form of kinetic energy, what is its velocity?

Exercise:

Problem:

What is the kinetic energy in MeV of a π-meson that lives 1.40×10^{-16} s as measured in the laboratory, and 0.840×10^{-16} s when at rest relative to an observer, given that its rest energy is 135 MeV?

Solution:

90.0 MeV

Exercise:

Problem:

Find the kinetic energy in MeV of a neutron with a measured life span of 2065 s, given its rest energy is 939.6 MeV, and rest life span is 900s.

Exercise:

Problem:

(a) Show that $(pc)^2/(mc^2)^2 = \gamma^2 - 1$. This means that at large velocities $pc >> mc^2$. (b) Is $E \approx$ pc when $\gamma = 30.0$, as for the astronaut discussed in the twin paradox?

Solution:

(a)
$$E^2 = p^2c^2 + m^2c^4 = \gamma^2 m^2 c^4\text{, so that}$$
$$p^2c^2 = \left(\gamma^2 - 1\right)m^2c^4 \text{ , and therefore}$$
$$\frac{(pc)^2}{\left(mc^2\right)^2} = \gamma^2 - 1$$

(b) yes

Exercise:

Problem:

One cosmic ray neutron has a velocity of $0.250c$ relative to the Earth. (a) What is the neutron's total energy in MeV? (b) Find its momentum. (c) Is $E \approx$ pc in this situation? Discuss in terms of the equation given in part (a) of the previous problem.

Exercise:

Problem:

What is γ for a proton having a mass energy of 938.3 MeV accelerated through an effective potential of 1.0 TV (teravolt) at Fermilab outside Chicago?

Solution:

1.07×10^3

Exercise:

Problem:

(a) What is the effective accelerating potential for electrons at the Stanford Linear Accelerator, if $\gamma = 1.00 \times 10^5$ for them? (b) What is their total energy (nearly the same as kinetic in this case) in GeV?

Exercise:

Problem:

(a) Using data from [link], find the mass destroyed when the energy in a barrel of crude oil is released. (b) Given these barrels contain 200 liters and assuming the density of crude oil is $750\ \mathrm{kg/m^3}$, what is the ratio of mass destroyed to original mass, $\Delta m/m$?

Solution:

$6.56\times 10^{-8}\ \mathrm{kg}$

4.37×10^{-10}

Exercise:

Problem:

(a) Calculate the energy released by the destruction of 1.00 kg of mass. (b) How many kilograms could be lifted to a 10.0 km height by this amount of energy?

Exercise:

Problem:

A Van de Graaff accelerator utilizes a 50.0 MV potential difference to accelerate charged particles such as protons. (a) What is the velocity of a proton accelerated by such a potential? (b) An electron?

Solution:

$0.314c$

$0.99995c$

Exercise:

Problem:

Suppose you use an average of $500\ \mathrm{kW{\cdot}h}$ of electric energy per month in your home. (a) How long would 1.00 g of mass converted to electric energy with an efficiency of 38.0% last you? (b) How many homes could be supplied at the $500\ \mathrm{kW{\cdot}h}$ per month rate for one year by the energy from the described mass conversion?

Exercise:

Problem:

(a) A nuclear power plant converts energy from nuclear fission into electricity with an efficiency of 35.0%. How much mass is destroyed in one year to produce a continuous 1000 MW of electric power? (b) Do you think it would be possible to observe this mass loss if the total mass of the fuel is 10^4 kg?

Solution:

(a) 1.00 kg

(b) This much mass would be measurable, but probably not observable just by looking because it is 0.01% of the total mass.

Exercise:

Problem:

Nuclear-powered rockets were researched for some years before safety concerns became paramount. (a) What fraction of a rocket's mass would have to be destroyed to get it into a low Earth orbit, neglecting the decrease in gravity? (Assume an orbital altitude of 250 km, and calculate both the kinetic energy (classical) and the gravitational potential energy needed.) (b) If the ship has a mass of 1.00×10^{5} kg (100 tons), what total yield nuclear explosion in tons of TNT is needed?

Exercise:

Problem:

The Sun produces energy at a rate of 4.00×10^{26} W by the fusion of hydrogen. (a) How many kilograms of hydrogen undergo fusion each second? (b) If the Sun is 90.0% hydrogen and half of this can undergo fusion before the Sun changes character, how long could it produce energy at its current rate? (c) How many kilograms of mass is the Sun losing per second? (d) What fraction of its mass will it have lost in the time found in part (b)?

Solution:

(a) $6.3 \times 10^{11}\ \mathrm{kg/s}$

(b) $4.5 \times 10^{10}\ \mathrm{y}$

(c) $4.44 \times 10^{9}\ \mathrm{kg}$

(d) 0.32%

Exercise:

Problem: **Unreasonable Results**

A proton has a mass of 1.67×10^{-27} kg. A physicist measures the proton's total energy to be 50.0 MeV. (a) What is the proton's kinetic energy? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: Construct Your Own Problem

Consider a highly relativistic particle. Discuss what is meant by the term “highly relativistic.” (Note that, in part, it means that the particle cannot be massless.) Construct a problem in which you calculate the wavelength of such a particle and show that it is very nearly the same as the wavelength of a massless particle, such as a photon, with the same energy. Among the things to be considered are the rest energy of the particle (it should be a known particle) and its total energy, which should be large compared to its rest energy.

Exercise:

Problem: Construct Your Own Problem

Consider an astronaut traveling to another star at a relativistic velocity. Construct a problem in which you calculate the time for the trip as observed on the Earth and as observed by the astronaut. Also calculate the amount of mass that must be converted to energy to get the astronaut and ship to the velocity travelled. Among the things to be considered are the distance to the star, the velocity, and the mass of the astronaut and ship. Unless your instructor directs you otherwise, do not include any energy given to other masses, such as rocket propellants.

Glossary

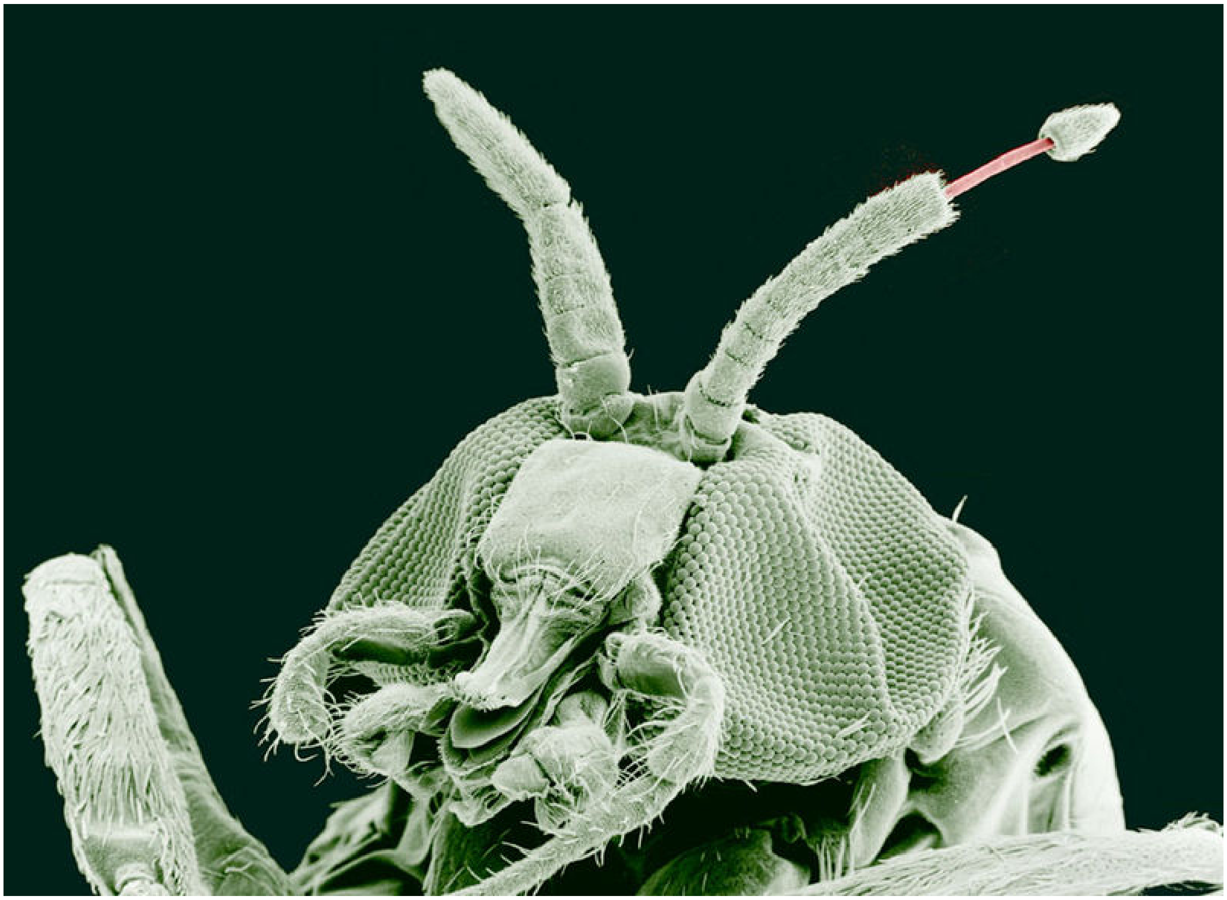
total energy
: defined as $E=\gamma mc^2$, where $\gamma=\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$

rest energy
: the energy stored in an object at rest: $E_0=mc^2$

relativistic kinetic energy
: the kinetic energy of an object moving at relativistic speeds: $\text{KE}_{\text{rel}}=(\gamma-1)\,mc^2$, where $\gamma=\frac{1}{\sqrt{1-\frac{v^2}{c^2}}}$

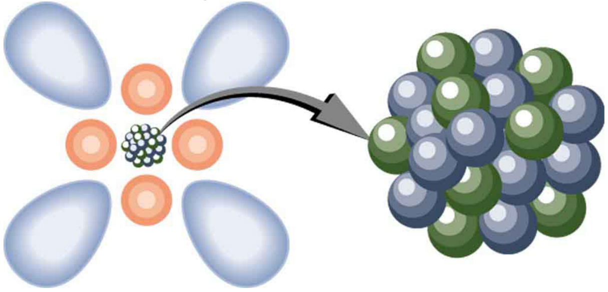
Introduction to Quantum Physics class="introduction"

A black fly imaged by an electron microscope is as monstrous as any science-fiction creature. (credit: U.S. Department of Agriculture via Wikimedia Commons)



Quantum mechanics is the branch of physics needed to deal with submicroscopic objects. Because these objects are smaller than we can observe directly with our senses and generally must be observed with the aid of instruments, parts of quantum mechanics seem as foreign and bizarre as parts of relativity. But, like relativity, quantum mechanics has been shown to be valid—truth is often stranger than fiction.

Certain aspects of quantum mechanics are familiar to us. We accept as fact that matter is composed of atoms, the smallest unit of an element, and that these atoms combine to form molecules, the smallest unit of a compound. (See [link].) While we cannot see the individual water molecules in a stream, for example, we are aware that this is because molecules are so small and so numerous in that stream. When introducing atoms, we commonly say that electrons orbit atoms in discrete shells around a tiny nucleus, itself composed of smaller particles called protons and neutrons. We are also aware that electric charge comes in tiny units carried almost entirely by electrons and protons. As with water molecules in a stream, we do not notice individual charges in the current through a lightbulb, because the charges are so small and so numerous in the macroscopic situations we sense directly.



Atoms and their substructure are familiar examples of objects that require quantum mechanics to be fully explained. Certain of their characteristics, such as the discrete electron shells, are classical physics explanations. In quantum mechanics we conceptualize discrete “electron clouds” around the nucleus.

Note:

Making Connections: Realms of Physics

Classical physics is a good approximation of modern physics under conditions first discussed in the The Nature of Science and Physics. Quantum mechanics is valid in general, and it must be used rather than classical physics to describe small objects, such as atoms.

Atoms, molecules, and fundamental electron and proton charges are all examples of physical entities that are **quantized**—that is, they appear only in certain discrete values and do not have every conceivable value.

Quantized is the opposite of continuous. We cannot have a fraction of an atom, or part of an electron’s charge, or 14-1/3 cents, for example. Rather, everything is built of integral multiples of these substructures. Quantum physics is the branch of physics that deals with small objects and the quantization of various entities, including energy and angular momentum. Just as with classical physics, quantum physics has several subfields, such as mechanics and the study of electromagnetic forces. The **correspondence principle** states that in the classical limit (large, slow-moving objects), **quantum mechanics** becomes the same as classical physics. In this chapter, we begin the development of quantum mechanics and its description of the strange submicroscopic world. In later chapters, we will examine many areas, such as atomic and nuclear physics, in which quantum mechanics is crucial.

Glossary

quantized
: the fact that certain physical entities exist only with particular discrete values and not every conceivable value

correspondence principle
: in the classical limit (large, slow-moving objects), quantum mechanics becomes the same as classical physics

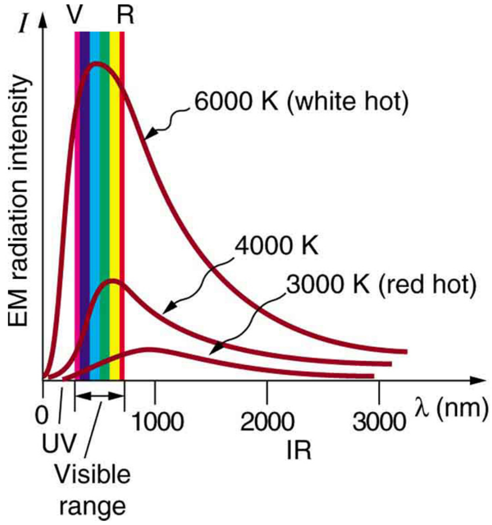
quantum mechanics
: the branch of physics that deals with small objects and with the quantization of various entities, especially energy

Quantization of Energy

- Explain Max Planck’s contribution to the development of quantum mechanics.
- Explain why atomic spectra indicate quantization.

Planck’s Contribution

Energy is quantized in some systems, meaning that the system can have only certain energies and not a continuum of energies, unlike the classical case. This would be like having only certain speeds at which a car can travel because its kinetic energy can have only certain values. We also find that some forms of energy transfer take place with discrete lumps of energy. While most of us are familiar with the quantization of matter into lumps called atoms, molecules, and the like, we are less aware that energy, too, can be quantized. Some of the earliest clues about the necessity of quantum mechanics over classical physics came from the quantization of energy.



Graphs of blackbody radiation (from an ideal radiator) at three different radiator temperatures. The intensity or rate of radiation emission increases dramatically with temperature, and the peak of the spectrum shifts toward the visible and ultraviolet parts of the spectrum. The shape of the spectrum cannot be described with classical physics.

Where is the quantization of energy observed? Let us begin by considering the emission and absorption of electromagnetic (EM) radiation. The EM spectrum radiated by a hot solid is linked directly to the solid’s temperature. (See [link].) An ideal radiator is one that has an emissivity of 1 at all wavelengths and, thus, is jet black. Ideal radiators are therefore called **blackbodies**, and their EM radiation is called **blackbody radiation**. It was discussed that the total intensity of the radiation varies as T^4, the fourth power of the absolute temperature of the body, and that the peak of the spectrum shifts to shorter wavelengths at higher temperatures. All of this seems quite continuous, but it was the curve of the spectrum of intensity versus wavelength that gave a clue that the energies of the atoms in the solid are quantized. In fact, providing a theoretical explanation for the experimentally measured shape of the spectrum was a mystery at the turn of the century. When this “ultraviolet catastrophe” was eventually solved, the answers led to new technologies such as computers and the sophisticated imaging techniques described in earlier chapters. Once again, physics as an enabling science changed the way we live.

The German physicist Max Planck (1858–1947) used the idea that atoms and molecules in a body act like oscillators to absorb and emit radiation. The energies of the oscillating atoms and molecules had to be quantized to correctly describe the shape of the blackbody spectrum. Planck deduced that the energy of an oscillator having a frequency f is given by

Equation:

$$E=\left(n+\frac{1}{2}\right)\text{hf}.$$

Here n is any nonnegative integer (0, 1, 2, 3, …). The symbol h stands for **Planck's constant**, given by

Equation:

$$h=6.626\times 10^{-34}\text{ J}\cdot\text{s}.$$

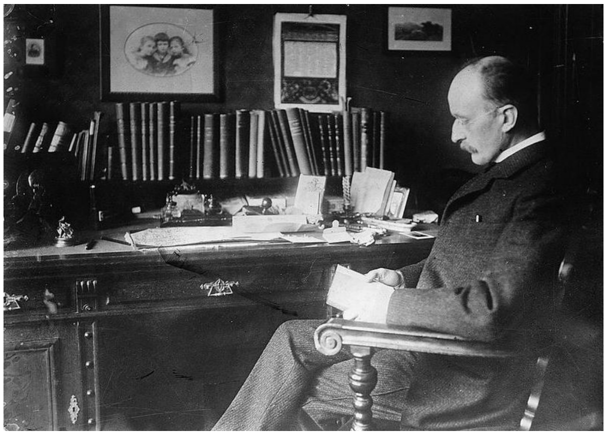
The equation $E=\left(n+\frac{1}{2}\right)\text{hf}$ means that an oscillator having a frequency f (emitting and absorbing EM radiation of frequency f) can have its energy increase or decrease only in *discrete* steps of size

Equation:

$$\Delta E=\text{hf}.$$

It might be helpful to mention some macroscopic analogies of this quantization of energy phenomena. This is like a pendulum that has a characteristic oscillation frequency but can swing with only certain amplitudes. Quantization of energy also resembles a standing wave on a string that allows only particular harmonics described by integers. It is also similar to going up and down a hill using discrete stair steps rather than being able to move up and down a continuous slope. Your potential energy takes on discrete values as you move from step to step.

Using the quantization of oscillators, Planck was able to correctly describe the experimentally known shape of the blackbody spectrum. This was the first indication that energy is sometimes quantized on a small scale and earned him the Nobel Prize in Physics in 1918. Although Planck's theory comes from observations of a macroscopic object, its analysis is based on atoms and molecules. It was such a revolutionary departure from classical physics that Planck himself was reluctant to accept his own idea that energy states are not continuous. The general acceptance of Planck's energy quantization was greatly enhanced by Einstein's explanation of the photoelectric effect (discussed in the next section), which took energy quantization a step further. Planck was fully involved in the development of both early quantum mechanics and relativity. He quickly embraced Einstein’s special relativity, published in 1905, and in 1906 Planck was the first to suggest the correct formula for relativistic momentum, $p = \gamma mu$.

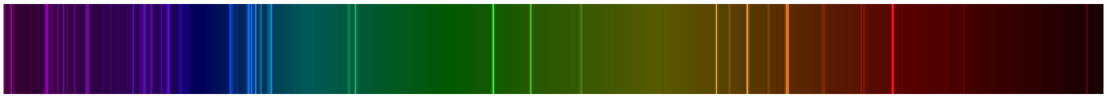


The German physicist Max Planck had a major influence on the early development of quantum mechanics, being the first to recognize that energy is sometimes quantized. Planck also made important contributions to special relativity and classical physics. (credit: Library of Congress, Prints and Photographs Division via Wikimedia Commons)

Note that Planck’s constant h is a very small number. So for an infrared frequency of 10^{14} Hz being emitted by a blackbody, for example, the difference between energy levels is only $\Delta E = \text{hf}=(6.63 \times 10^{-34}\ \text{J}\cdot\text{s})(10^{14}\ \text{Hz})= 6.63 \times 10^{-20}\ \text{J}$, or about 0.4 eV. This 0.4 eV of energy is significant compared with typical atomic energies, which are on the order of an electron volt, or thermal energies, which are typically fractions of an electron volt. But on a macroscopic or classical scale, energies are typically on the order of joules. Even if macroscopic energies are quantized, the quantum steps are too small to be noticed. This is an example of the correspondence principle. For a large object, quantum mechanics produces results indistinguishable from those of classical physics.

Atomic Spectra

Now let us turn our attention to the *emission and absorption of EM radiation by gases*. The Sun is the most common example of a body containing gases emitting an EM spectrum that includes visible light. We also see examples in neon signs and candle flames. Studies of emissions of hot gases began more than two centuries ago, and it was soon recognized that these emission spectra contained huge amounts of information. The type of gas and its temperature, for example, could be determined. We now know that these EM emissions come from electrons transitioning between energy levels in individual atoms and molecules; thus, they are called **atomic spectra**. Atomic spectra remain an important analytical tool today. [link] shows an example of an emission spectrum obtained by passing an electric discharge through a material. One of the most important characteristics of these spectra is that they are discrete. By this we mean that only certain wavelengths, and hence frequencies, are emitted. This is called a line spectrum. If frequency and energy are associated as $\Delta E = \text{hf}$, the energies of the electrons in the emitting atoms and molecules are quantized. This is discussed in more detail later in this chapter.



Emission spectrum of oxygen. When an electrical discharge is passed through a substance, its atoms and molecules absorb energy, which is reemitted as EM radiation. The discrete nature of these emissions implies that the energy states of the atoms and molecules are quantized. Such atomic spectra were used as analytical tools for many decades before it was understood why they are quantized. (credit: Teravolt, Wikimedia Commons)

It was a major puzzle that atomic spectra are quantized. Some of the best minds of 19th-century science failed to explain why this might be. Not until the second decade of the 20th century did an answer based on quantum mechanics begin to emerge. Again a macroscopic or classical body of gas was involved in the studies, but the effect, as we shall see, is due to individual atoms and molecules.

Note:
PhET Explorations: Models of the Hydrogen Atom
How did scientists figure out the structure of atoms without looking at them? Try out different models by shooting light at the atom. Check how the prediction of the model matches the experimental results.

https://archive.cnx.org/specials/d77cc1d0-33e4-11e6-b016-6726afecd2be/hydrogen-atom/#sim-hydrogen-atom

Section Summary

- The first indication that energy is sometimes quantized came from blackbody radiation, which is the emission of EM radiation by an object with an emissivity of 1.
- Planck recognized that the energy levels of the emitting atoms and molecules were quantized, with only the allowed values of $E = \left(n + \frac{1}{2}\right)\text{hf}$, where n is any non-negative integer (0, 1, 2, 3, …).
- h is Planck’s constant, whose value is $h = 6.626 \times 10^{–34}\ \text{J} \cdot \text{s}$.
- Thus, the oscillatory absorption and emission energies of atoms and molecules in a blackbody could increase or decrease only in steps of

size $\Delta E = \text{hf}$ where f is the frequency of the oscillatory nature of the absorption and emission of EM radiation.
- Another indication of energy levels being quantized in atoms and molecules comes from the lines in atomic spectra, which are the EM emissions of individual atoms and molecules.

Conceptual Questions

Exercise:

Problem:

Give an example of a physical entity that is quantized. State specifically what the entity is and what the limits are on its values.

Exercise:

Problem:

Give an example of a physical entity that is not quantized, in that it is continuous and may have a continuous range of values.

Exercise:

Problem:

What aspect of the blackbody spectrum forced Planck to propose quantization of energy levels in its atoms and molecules?

Exercise:

Problem:

If Planck’s constant were large, say 10^{34} times greater than it is, we would observe macroscopic entities to be quantized. Describe the motions of a child’s swing under such circumstances.

Exercise:

Problem: Why don’t we notice quantization in everyday events?

Problems & Exercises

Exercise:

Problem:

A LiBr molecule oscillates with a frequency of $1.7 \times 10^{13}\ \text{Hz}$. (a) What is the difference in energy in eV between allowed oscillator states? (b) What is the approximate value of n for a state having an energy of 1.0 eV?

Solution:

(a) 0.070 eV

(b) 14

Exercise:

Problem:

The difference in energy between allowed oscillator states in HBr molecules is 0.330 eV. What is the oscillation frequency of this molecule?

Exercise:

Problem:

A physicist is watching a 15-kg orangutan at a zoo swing lazily in a tire at the end of a rope. He (the physicist) notices that each oscillation takes 3.00 s and hypothesizes that the energy is quantized. (a) What is the difference in energy in joules between allowed oscillator states? (b) What is the value of n for a state where the energy is 5.00 J? (c) Can the quantization be observed?

Solution:

(a) $2.21 \times 10^{34}\ \text{J}$

(b) 2.26×10^{34}

(c) No

Glossary

blackbody
an ideal radiator, which can radiate equally well at all wavelengths

blackbody radiation
the electromagnetic radiation from a blackbody

Planck’s constant
$h = 6.626 \times 10^{-34}\ \mathrm{J \cdot s}$

atomic spectra
the electromagnetic emission from atoms and molecules

The Photoelectric Effect

- Describe a typical photoelectric-effect experiment.
- Determine the maximum kinetic energy of photoelectrons ejected by photons of one energy or wavelength, when given the maximum kinetic energy of photoelectrons for a different photon energy or wavelength.

When light strikes materials, it can eject electrons from them. This is called the **photoelectric effect**, meaning that light (*photo*) produces electricity. One common use of the photoelectric effect is in light meters, such as those that adjust the automatic iris on various types of cameras. In a similar way, another use is in solar cells, as you probably have in your calculator or have seen on a roof top or a roadside sign. These make use of the photoelectric effect to convert light into electricity for running different devices.

The photoelectric effect can be observed by allowing light to fall on the metal plate in this evacuated tube. Electrons ejected by the light are collected on the collector wire and measured as a current. A retarding voltage between the collector wire and plate can then be adjusted so as to determine the energy of the ejected electrons. For example, if it is sufficiently negative, no electrons will reach the wire. (credit: P.P. Urone)

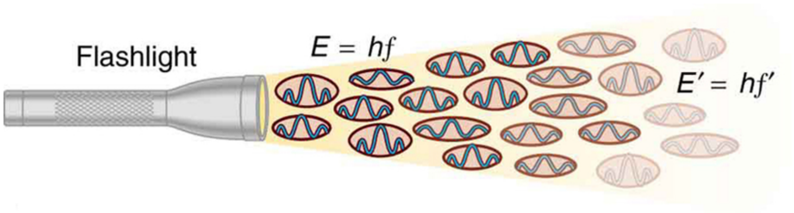
This effect has been known for more than a century and can be studied using a device such as that shown in [link]. This figure shows an evacuated tube with a metal plate and a collector wire that are connected by a variable voltage source, with the collector more negative than the plate. When light (or other EM radiation) strikes the plate in the evacuated tube, it may eject electrons. If the electrons have energy in electron volts (eV) greater than the potential difference between the plate and the wire in volts, some electrons will be collected on the wire. Since the electron energy in eV is qV, where q is the electron charge and V is the potential difference, the electron energy can be measured by adjusting the retarding voltage between the wire and the plate. The voltage that stops the electrons from reaching the wire equals the energy in eV. For example, if $–3.00\text{ V}$ barely stops the electrons, their energy is 3.00 eV. The number of electrons ejected can be determined by measuring the current between the wire and plate. The more light, the more electrons; a little circuitry allows this device to be used as a light meter.

What is really important about the photoelectric effect is what Albert Einstein deduced from it. Einstein realized that there were several characteristics of the photoelectric effect that could be explained only if *EM radiation is itself quantized*: the apparently continuous stream of energy in an EM wave is actually composed of energy quanta called photons. In his explanation of the photoelectric effect, Einstein defined a quantized unit or quantum of EM energy, which we now call a **photon**, with an energy proportional to the frequency of EM radiation. In equation form, the **photon energy** is

Equation:

$$E=\text{hf},$$

where E is the energy of a photon of frequency f and h is Planck's constant. This revolutionary idea looks similar to Planck's quantization of energy states in blackbody oscillators, but it is quite different. It is the quantization of EM radiation itself. EM waves are composed of photons and are not continuous smooth waves as described in previous chapters on optics. Their energy is absorbed and emitted in lumps, not continuously. This is exactly consistent with Planck's quantization of energy levels in blackbody oscillators, since these oscillators increase and decrease their energy in steps of hf by absorbing and emitting photons having $E=\text{hf}$. We do not observe this with our eyes, because there are so many photons in common light sources that individual photons go unnoticed. (See [link].) The next section of the text (Photon Energies and the Electromagnetic Spectrum) is devoted to a discussion of photons and some of their characteristics and implications. For now, we will use the photon concept to explain the photoelectric effect, much as Einstein did.



An EM wave of frequency f is composed of photons, or individual quanta of EM radiation. The energy of each photon is $E = \text{hf}$, where h is Planck’s constant and f is the frequency of the EM radiation. Higher intensity means more photons per unit area. The flashlight emits large numbers of photons of many different frequencies, hence others have energy $E\prime = \text{hf}\prime$, and so on.

The photoelectric effect has the properties discussed below. All these properties are consistent with the idea that individual photons of EM radiation are absorbed by individual electrons in a material, with the electron gaining the photon’s energy. Some of these properties are inconsistent with the idea that EM radiation is a simple wave. For simplicity, let us consider what happens with monochromatic EM radiation in which all photons have the same energy hf.

1. If we vary the frequency of the EM radiation falling on a material, we find the following: For a given material, there is a threshold frequency f_0 for the EM radiation below which no electrons are ejected, regardless of intensity. Individual photons interact with individual electrons. Thus if the photon energy is too small to break an electron away, no electrons will be ejected. If EM radiation was a simple wave, sufficient energy could be obtained by increasing the intensity.
2. *Once EM radiation falls on a material, electrons are ejected without delay*. As soon as an individual photon of a sufficiently high frequency is absorbed by an individual electron, the electron is ejected. If the EM

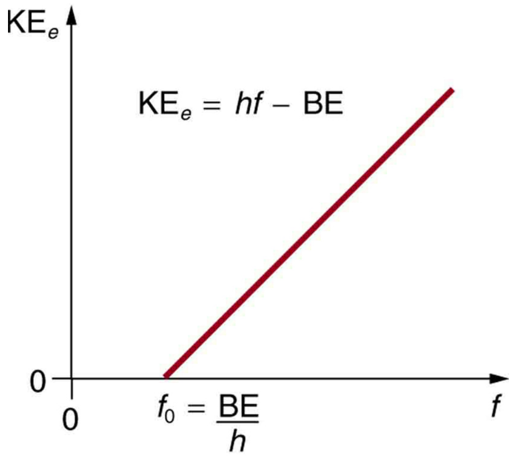
radiation were a simple wave, several minutes would be required for sufficient energy to be deposited to the metal surface to eject an electron.

3. The number of electrons ejected per unit time is proportional to the intensity of the EM radiation and to no other characteristic. High-intensity EM radiation consists of large numbers of photons per unit area, with all photons having the same characteristic energy hf.
4. If we vary the intensity of the EM radiation and measure the energy of ejected electrons, we find the following: *The maximum kinetic energy of ejected electrons is independent of the intensity of the EM radiation*. Since there are so many electrons in a material, it is extremely unlikely that two photons will interact with the same electron at the same time, thereby increasing the energy given it. Instead (as noted in 3 above), increased intensity results in more electrons of the same energy being ejected. If EM radiation were a simple wave, a higher intensity could give more energy, and higher-energy electrons would be ejected.
5. The kinetic energy of an ejected electron equals the photon energy minus the binding energy of the electron in the specific material. An individual photon can give all of its energy to an electron. The photon’s energy is partly used to break the electron away from the material. The remainder goes into the ejected electron’s kinetic energy. In equation form, this is given by

   Equation:

   $$\text{KE}_e = \text{hf} - \text{BE},$$

   where KE_e is the maximum kinetic energy of the ejected electron, hf is the photon’s energy, and BE is the **binding energy** of the electron to the particular material. (BE is sometimes called the *work function* of the material.) This equation, due to Einstein in 1905, explains the properties of the photoelectric effect quantitatively. An individual photon of EM radiation (it does not come any other way) interacts with an individual electron, supplying enough energy, BE, to break it away, with the remainder going to kinetic energy. The binding energy is $\text{BE} = hf_0$, where f_0 is the threshold frequency for the particular material. [link] shows a graph of maximum KE_e versus the frequency of incident EM radiation falling on a particular material.



Photoelectric effect. A graph of the kinetic energy of an ejected electron, KE_e, versus the frequency of EM radiation impinging on a certain material. There is a threshold frequency below which no electrons are ejected, because the individual photon interacting with an individual electron has insufficient energy to break it away. Above the threshold energy, KE_e increases linearly with f, consistent with $\mathrm{KE}_e = \mathrm{hf} - \mathrm{BE}$. The slope of this line is h — the data can be used to determine Planck’s constant experimentally. Einstein gave the first successful explanation of such data by proposing

the idea of photons—
quanta of EM radiation.

Einstein’s idea that EM radiation is quantized was crucial to the beginnings of quantum mechanics. It is a far more general concept than its explanation of the photoelectric effect might imply. All EM radiation can also be modeled in the form of photons, and the characteristics of EM radiation are entirely consistent with this fact. (As we will see in the next section, many aspects of EM radiation, such as the hazards of ultraviolet (UV) radiation, can be explained *only* by photon properties.) More famous for modern relativity, Einstein planted an important seed for quantum mechanics in 1905, the same year he published his first paper on special relativity. His explanation of the photoelectric effect was the basis for the Nobel Prize awarded to him in 1921. Although his other contributions to theoretical physics were also noted in that award, special and general relativity were not fully recognized in spite of having been partially verified by experiment by 1921. Although hero-worshipped, this great man never received Nobel recognition for his most famous work—relativity.

Example:
Calculating Photon Energy and the Photoelectric Effect: A Violet Light

(a) What is the energy in joules and electron volts of a photon of 420-nm violet light? (b) What is the maximum kinetic energy of electrons ejected from calcium by 420-nm violet light, given that the binding energy (or work function) of electrons for calcium metal is 2.71 eV?

Strategy

To solve part (a), note that the energy of a photon is given by $E = \text{hf}$. For part (b), once the energy of the photon is calculated, it is a straightforward application of $\text{KE}_e = \text{hf}–\text{BE}$ to find the ejected electron’s maximum kinetic energy, since BE is given.

Solution for (a)

Photon energy is given by

Equation:

$$E = \mathrm{hf}$$

Since we are given the wavelength rather than the frequency, we solve the familiar relationship $c = f\lambda$ for the frequency, yielding

Equation:

$$f = \frac{c}{\lambda}.$$

Combining these two equations gives the useful relationship

Equation:

$$E = \frac{\mathrm{hc}}{\lambda}.$$

Now substituting known values yields

Equation:

$$E = \frac{\left(6.63 \times 10^{–34}\ \mathrm{J} \cdot \mathrm{s}\right)\left(3.00 \times 10^{8}\ \mathrm{m/s}\right)}{420 \times 10^{–9}\ \mathrm{m}} = 4.74 \times 10^{–19}\ \mathrm{J}.$$

Converting to eV, the energy of the photon is

Equation:

$$E = \left(4.74 \times 10^{–19}\ \mathrm{J}\right)\frac{1\ \mathrm{eV}}{1.6 \times 10^{–19}\ \mathrm{J}} = 2.96\ \mathrm{eV}.$$

Solution for (b)

Finding the kinetic energy of the ejected electron is now a simple application of the equation $\mathrm{KE}_{e} = \mathrm{hf}–\mathrm{BE}$. Substituting the photon energy and binding energy yields

Equation:

$$\mathrm{KE}_{e} = \mathrm{hf}– \mathrm{BE} = 2.96\ \mathrm{eV} – 2.71\ \mathrm{eV} = 0.246\ \mathrm{eV}.$$

Discussion

The energy of this 420-nm photon of violet light is a tiny fraction of a joule, and so it is no wonder that a single photon would be difficult for us to sense directly—humans are more attuned to energies on the order of joules. But looking at the energy in electron volts, we can see that this photon has enough energy to affect atoms and molecules. A DNA molecule can be broken with about 1 eV of energy, for example, and typical atomic and molecular energies are on the order of eV, so that the UV photon in this example could have biological effects. The ejected electron (called a *photoelectron*) has a rather low energy, and it would not travel far, except in a vacuum. The electron would be stopped by a retarding potential of but 0.26 eV. In fact, if the photon wavelength were longer and its energy less than 2.71 eV, then the formula would give a negative kinetic energy, an impossibility. This simply means that the 420-nm photons with their 2.96-eV energy are not much above the frequency threshold. You can show for yourself that the threshold wavelength is 459 nm (blue light). This means that if calcium metal is used in a light meter, the meter will be insensitive to wavelengths longer than those of blue light. Such a light meter would be completely insensitive to red light, for example.

Note:
PhET Explorations: Photoelectric Effect
See how light knocks electrons off a metal target, and recreate the experiment that spawned the field of quantum mechanics.

https://archive.cnx.org/specials/cf1152da-eae8-11e5-b874-f779884a9994/photoelectric-effect/#sim-photoelectric-effect

Section Summary

- The photoelectric effect is the process in which EM radiation ejects electrons from a material.
- Einstein proposed photons to be quanta of EM radiation having energy $E = \text{hf}$, where f is the frequency of the radiation.

- All EM radiation is composed of photons. As Einstein explained, all characteristics of the photoelectric effect are due to the interaction of individual photons with individual electrons.
- The maximum kinetic energy KE_e of ejected electrons (photoelectrons) is given by $\text{KE}_e = \text{hf} - \text{BE}$, where hf is the photon energy and BE is the binding energy (or work function) of the electron to the particular material.

Conceptual Questions

Exercise:

Problem:

Is visible light the only type of EM radiation that can cause the photoelectric effect?

Exercise:

Problem:

Which aspects of the photoelectric effect cannot be explained without photons? Which can be explained without photons? Are the latter inconsistent with the existence of photons?

Exercise:

Problem:

Is the photoelectric effect a direct consequence of the wave character of EM radiation or of the particle character of EM radiation? Explain briefly.

Exercise:

Problem:

Insulators (nonmetals) have a higher BE than metals, and it is more difficult for photons to eject electrons from insulators. Discuss how this relates to the free charges in metals that make them good conductors.

Exercise:

Problem:

If you pick up and shake a piece of metal that has electrons in it free to move as a current, no electrons fall out. Yet if you heat the metal, electrons can be boiled off. Explain both of these facts as they relate to the amount and distribution of energy involved with shaking the object as compared with heating it.

Problems & Exercises

Exercise:

Problem:

What is the longest-wavelength EM radiation that can eject a photoelectron from silver, given that the binding energy is 4.73 eV? Is this in the visible range?

Solution:

263 nm

Exercise:

Problem:

Find the longest-wavelength photon that can eject an electron from potassium, given that the binding energy is 2.24 eV. Is this visible EM radiation?

Exercise:

Problem:

What is the binding energy in eV of electrons in magnesium, if the longest-wavelength photon that can eject electrons is 337 nm?

Solution:

3.69 eV

Exercise:

Problem:

Calculate the binding energy in eV of electrons in aluminum, if the longest-wavelength photon that can eject them is 304 nm.

Exercise:

Problem:

What is the maximum kinetic energy in eV of electrons ejected from sodium metal by 450-nm EM radiation, given that the binding energy is 2.28 eV?

Solution:

0.483 eV

Exercise:

Problem:

UV radiation having a wavelength of 120 nm falls on gold metal, to which electrons are bound by 4.82 eV. What is the maximum kinetic energy of the ejected photoelectrons?

Exercise:

Problem:

Violet light of wavelength 400 nm ejects electrons with a maximum kinetic energy of 0.860 eV from sodium metal. What is the binding energy of electrons to sodium metal?

Solution:

2.25 eV

Exercise:

Problem:

UV radiation having a 300-nm wavelength falls on uranium metal, ejecting 0.500-eV electrons. What is the binding energy of electrons to uranium metal?

Exercise:

Problem:

What is the wavelength of EM radiation that ejects 2.00-eV electrons from calcium metal, given that the binding energy is 2.71 eV? What type of EM radiation is this?

Solution:

(a) 264 nm

(b) Ultraviolet

Exercise:

Problem:

Find the wavelength of photons that eject 0.100-eV electrons from potassium, given that the binding energy is 2.24 eV. Are these photons visible?

Exercise:

Problem:

What is the maximum velocity of electrons ejected from a material by 80-nm photons, if they are bound to the material by 4.73 eV?

Solution:

$1.95 \times 10^{6}\ \mathrm{m/s}$

Exercise:

Problem:

Photoelectrons from a material with a binding energy of 2.71 eV are ejected by 420-nm photons. Once ejected, how long does it take these electrons to travel 2.50 cm to a detection device?

Exercise:

Problem:

A laser with a power output of 2.00 mW at a wavelength of 400 nm is projected onto calcium metal. (a) How many electrons per second are ejected? (b) What power is carried away by the electrons, given that the binding energy is 2.71 eV?

Solution:

(a) $4.02 \times 10^{15}\ /\mathrm{s}$

(b) 0.256 mW

Exercise:

Problem:

(a) Calculate the number of photoelectrons per second ejected from a 1.00-mm ^{2} area of sodium metal by 500-nm EM radiation having an intensity of $1.30\ \mathrm{kW/m}^2$ (the intensity of sunlight above the Earth’s atmosphere). (b) Given that the binding energy is 2.28 eV, what power is carried away by the electrons? (c) The electrons carry away less power than brought in by the photons. Where does the other power go? How can it be recovered?

Exercise:

Problem: Unreasonable Results

Red light having a wavelength of 700 nm is projected onto magnesium metal to which electrons are bound by 3.68 eV. (a) Use $\mathrm{KE}_e = \mathrm{hf}{-}\mathrm{BE}$ to calculate the kinetic energy of the ejected electrons.

(b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) –1.90 eV

(b) Negative kinetic energy

(c) That the electrons would be knocked free.

Exercise:

Problem: Unreasonable Results

(a) What is the binding energy of electrons to a material from which 4.00-eV electrons are ejected by 400-nm EM radiation? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Glossary

photoelectric effect
the phenomenon whereby some materials eject electrons when light is shined on them

photon
a quantum, or particle, of electromagnetic radiation

photon energy
the amount of energy a photon has; $E = \text{hf}$

binding energy
also called the *work function*; the amount of energy necessary to eject an electron from a material

Photon Energies and the Electromagnetic Spectrum

- Explain the relationship between the energy of a photon in joules or electron volts and its wavelength or frequency.
- Calculate the number of photons per second emitted by a monochromatic source of specific wavelength and power.

Ionizing Radiation

A photon is a quantum of EM radiation. Its energy is given by $E=\text{hf}$ and is related to the frequency f and wavelength λ of the radiation by

Equation:

$$E=\text{hf}=\frac{\text{hc}}{\lambda}\text{(energy of a photon)},$$

where E is the energy of a single photon and c is the speed of light. When working with small systems, energy in eV is often useful. Note that Planck’s constant in these units is

Equation:

$$h=4\text{.}\text{14}\times {\text{10}}^{\text{–15}}\phantom{\rule{0.25em}{0ex}}\text{eV}\cdot \text{s.}$$

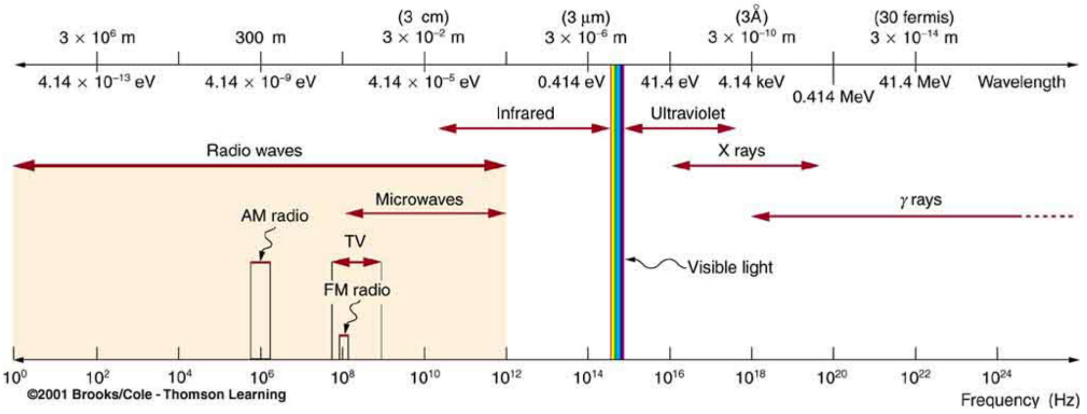
Since many wavelengths are stated in nanometers (nm), it is also useful to know that

Equation:

$$\text{hc}=\text{1240 eV}\cdot \text{nm.}$$

These will make many calculations a little easier.

All EM radiation is composed of photons. [link] shows various divisions of the EM spectrum plotted against wavelength, frequency, and photon energy. Previously in this book, photon characteristics were alluded to in the discussion of some of the characteristics of UV, x rays, and γ rays, the first of which start with frequencies just above violet in the visible spectrum. It was noted that these types of EM radiation have characteristics much different than visible light. We can now see that such properties arise because photon energy is larger at high frequencies.



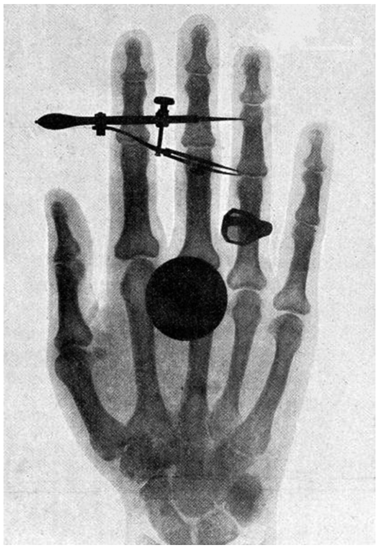
The EM spectrum, showing major categories as a function of photon energy in eV, as well as wavelength and frequency. Certain characteristics of EM radiation are directly attributable to photon energy alone.

Rotational energies of molecules	10^{-5} eV
Vibrational energies of molecules	0.1 eV
Energy between outer electron shells in atoms	1 eV
Binding energy of a weakly bound molecule	1 eV
Energy of red light	2 eV
Binding energy of a tightly bound molecule	10 eV
Energy to ionize atom or molecule	10 to 1000 eV

Representative Energies for Submicroscopic Effects (Order of Magnitude Only)

Photons act as individual quanta and interact with individual electrons, atoms, molecules, and so on. The energy a photon carries is, thus, crucial to the effects it has. [link] lists representative submicroscopic energies in eV. When we compare photon energies from the EM spectrum in [link] with energies in the table, we can see how effects vary with the type of EM radiation.

Gamma rays, a form of nuclear and cosmic EM radiation, can have the highest frequencies and, hence, the highest photon energies in the EM spectrum. For example, a γ-ray photon with $f = 10^{21}$ Hz has an energy $E = \text{hf} = 6.63 \times 10^{-13}\text{ J} = 4.14\text{ MeV}$. This is sufficient energy to ionize thousands of atoms and molecules, since only 10 to 1000 eV are needed per ionization. In fact, γ rays are one type of **ionizing radiation**, as are x rays and UV, because they produce ionization in materials that absorb them. Because so much ionization can be produced, a single γ-ray photon can cause significant damage to biological tissue, killing cells or damaging their ability to properly reproduce. When cell reproduction is disrupted, the result can be cancer, one of the known effects of exposure to ionizing radiation. Since cancer cells are rapidly reproducing, they are exceptionally sensitive to the disruption produced by ionizing radiation. This means that ionizing radiation has positive uses in cancer treatment as well as risks in producing cancer.



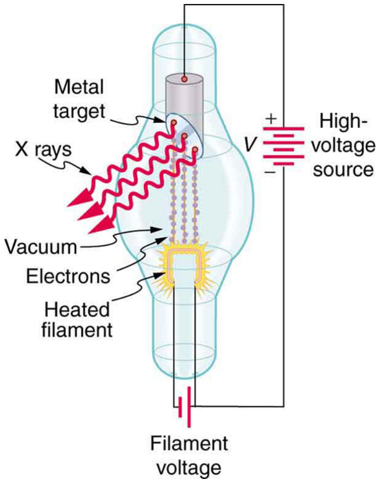
One of the first x-ray images, taken by Röentgen himself. The hand belongs to Bertha Röentgen, his wife. (credit: Wilhelm Conrad Röntgen, via Wikimedia Commons)

High photon energy also enables γ rays to penetrate materials, since a collision with a single atom or molecule is unlikely to absorb all the γ ray's energy. This can make γ rays useful as a probe, and they are sometimes used in medical imaging. **x rays**, as you can see in [link], overlap with the low-frequency end of the γ ray range. Since x rays have energies of keV and up, individual x-ray photons also can produce large amounts of ionization. At lower photon energies, x rays are not as penetrating as γ rays and are slightly less hazardous. X rays are ideal for medical imaging, their most common use, and a fact that was recognized immediately upon their discovery in 1895 by the German physicist W. C. Roentgen (1845–1923). (See [link].) Within one year of their discovery, x rays (for a time called Roentgen rays) were used for medical diagnostics. Roentgen received the 1901 Nobel Prize for the discovery of x rays.

Note:

Connections: Conservation of Energy

Once again, we find that conservation of energy allows us to consider the initial and final forms that energy takes, without having to make detailed calculations of the intermediate steps. [link] is solved by considering only the initial and final forms of energy.



X rays are produced when energetic electrons strike the copper anode of this cathode ray tube (CRT). Electrons (shown here as separate particles) interact individually with the material they strike, sometimes producing photons of EM radiation.

While γ rays originate in nuclear decay, x rays are produced by the process shown in [link]. Electrons ejected by thermal agitation from a hot filament in a vacuum tube are accelerated through a high voltage, gaining kinetic energy from the electrical potential energy. When they strike the anode, the electrons convert their kinetic energy to a variety of forms, including thermal energy. But since an accelerated charge radiates EM waves, and since the electrons act individually, photons are also produced. Some of these x-ray photons obtain the kinetic energy of the electron. The accelerated electrons originate at the cathode, so such a tube is called a cathode ray tube (CRT), and various versions of them are found in older TV and computer screens as well as in x-ray machines.

Example:
X-ray Photon Energy and X-ray Tube Voltage

Find the maximum energy in eV of an x-ray photon produced by electrons accelerated through a potential difference of 50.0 kV in a CRT like the one in [link].

Strategy

Electrons can give all of their kinetic energy to a single photon when they strike the anode of a CRT. (This is something like the photoelectric effect in reverse.) The kinetic energy of the electron comes from electrical potential energy. Thus we can simply equate the maximum photon energy to the electrical potential energy—that is, $hf = qV$. (We do not have to calculate each step from beginning to end if we know that all of the starting energy qV is converted to the final form hf.)

Solution

The maximum photon energy is $\text{hf} = \text{qV}$, where q is the charge of the electron and V is the accelerating voltage. Thus,

Equation:

$$\text{hf} = (1.60 \times 10^{-19}\text{ C})(50.0 \times 10^{3}\text{ V}).$$

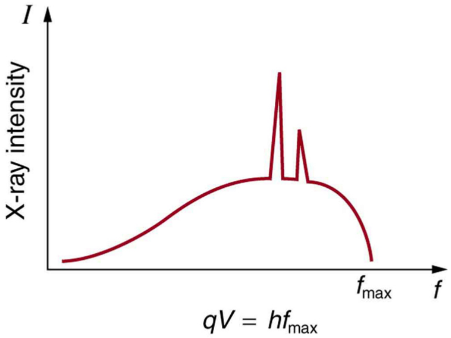
From the definition of the electron volt, we know $1\text{ eV} = 1.60 \times 10^{-19}\text{ J}$, where $1\text{ J} = 1\text{ C} \cdot \text{V}$. Gathering factors and converting energy to eV yields

Equation:

$$\text{hf} = (50.0 \times 10^{3})(1.60 \times 10^{-19}\text{ C} \cdot \text{V})\left(\frac{1\text{ eV}}{1.60 \times 10^{-19}\text{ C} \cdot \text{V}}\right) = (50.0 \times 10^{3})(1\text{ eV}) = 50.0\text{ keV}.$$

Discussion

This example produces a result that can be applied to many similar situations. If you accelerate a single elementary charge, like that of an electron, through a potential given in volts, then its energy in eV has the same numerical value. Thus a 50.0-kV potential generates 50.0 keV electrons, which in turn can produce photons with a maximum energy of 50 keV. Similarly, a 100-kV potential in an x-ray tube can generate up to 100-keV x-ray photons. Many x-ray tubes have adjustable voltages so that various energy x rays with differing energies, and therefore differing abilities to penetrate, can be generated.



X-ray spectrum obtained when energetic electrons strike a material. The smooth part of the spectrum is bremsstrahlung, while the peaks are characteristic of the anode material. Both are atomic processes that produce energetic photons known as x-ray photons.

[link] shows the spectrum of x rays obtained from an x-ray tube. There are two distinct features to the spectrum. First, the smooth distribution results from electrons being decelerated in the anode material. A curve like this is obtained by detecting many photons, and it is apparent that the maximum energy is unlikely. This decelerating process produces radiation that is called **bremsstrahlung** (German for *braking radiation*). The second feature is the existence of sharp peaks in the spectrum; these are called **characteristic x rays**, since they are characteristic of the anode material. Characteristic x rays come from atomic excitations unique to a given type of anode material. They are akin to lines in atomic spectra, implying the energy levels of atoms are quantized. Phenomena such as discrete atomic spectra and characteristic x rays are explored further in Atomic Physics.

Ultraviolet radiation (approximately 4 eV to 300 eV) overlaps with the low end of the energy range of x rays, but UV is typically lower in energy. UV comes from the de-excitation of atoms that may be part of a hot solid or gas. These atoms can be given energy that they later release as UV by numerous processes, including electric discharge, nuclear explosion, thermal agitation, and exposure to x rays. A UV photon has sufficient energy to ionize atoms and molecules, which makes its effects different from those of visible light. UV thus has some of the same biological effects as γ rays and x rays. For example, it can cause skin cancer and is used as a sterilizer. The major difference is that several UV photons are required to disrupt cell reproduction or kill a bacterium, whereas single γ-ray and X-ray photons can do the same damage. But since UV does have the energy to alter molecules, it can do what visible light cannot. One of the beneficial aspects of UV is that it triggers the production of vitamin D in the skin, whereas visible light has insufficient energy per photon to alter the molecules that trigger this production. Infantile jaundice is treated by exposing the baby to UV (with eye protection), called phototherapy, the beneficial effects of which are thought to be related to its ability to help prevent the buildup of potentially toxic bilirubin in the blood.

Example:
Photon Energy and Effects for UV

Short-wavelength UV is sometimes called vacuum UV, because it is strongly absorbed by air and must be studied in a vacuum. Calculate the photon energy in eV for 100-nm vacuum UV, and estimate the number of molecules it could ionize or break apart.

Strategy

Using the equation $E = \text{hf}$ and appropriate constants, we can find the photon energy and compare it with energy information in [link].

Solution

The energy of a photon is given by

Equation:

$$E = \text{hf} = \frac{\text{hc}}{\lambda}.$$

Using $\text{hc} = 1240\text{ eV} \cdot \text{nm}$, we find that

Equation:

$$E = \frac{\text{hc}}{\lambda} = \frac{1240\text{ eV} \cdot \text{nm}}{100\text{ nm}} = 12.4\text{ eV}.$$

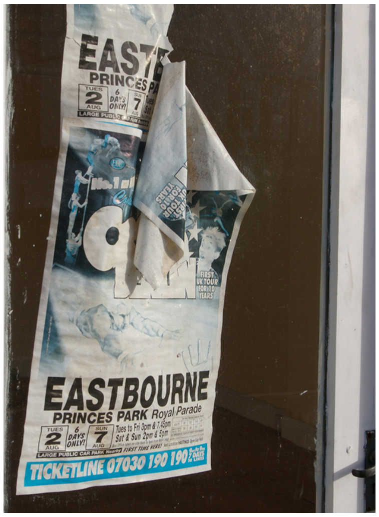
Discussion

According to [link], this photon energy might be able to ionize an atom or molecule, and it is about what is needed to break up a tightly bound molecule, since they are bound by approximately 10 eV. This photon energy could destroy about a dozen weakly bound molecules. Because of its high photon energy, UV disrupts atoms and molecules it interacts with. One good consequence is that all but the longest-wavelength UV is strongly absorbed and is easily blocked by sunglasses. In fact, most of the Sun's UV is absorbed by a thin layer of ozone in the upper atmosphere, protecting sensitive organisms on Earth. Damage to our ozone layer by the addition of such chemicals as CFC's has reduced this protection for us.

Visible Light

The range of photon energies for **visible light** from red to violet is 1.63 to 3.26 eV, respectively (left for this chapter's Problems and Exercises to verify). These energies are on the order of those between outer electron shells in atoms and molecules. This means that these photons can be absorbed by atoms and molecules. A *single* photon can actually stimulate the retina, for example, by altering a receptor molecule that then triggers a nerve impulse. Photons can be absorbed or emitted only by atoms and molecules that have precisely the correct quantized energy step to do so. For example, if a red photon of frequency f encounters a molecule that has an energy step, ΔE, equal to hf, then the photon can be absorbed. Violet flowers absorb red and reflect violet; this implies there is no energy step between levels in the receptor molecule equal to the violet photon's energy, but there is an energy step for the red.

There are some noticeable differences in the characteristics of light between the two ends of the visible spectrum that are due to photon energies. Red light has insufficient photon energy to expose most black-and-white film, and it is thus used to illuminate darkrooms where such film is developed. Since violet light has a higher photon energy, dyes that absorb violet tend to fade more quickly than those that do not. (See [link].) Take a look at some faded color posters in a storefront some time, and you will notice that the blues and violets are the last to fade. This is because other dyes, such as red and green dyes, absorb blue and violet photons, the higher energies of which break up their weakly bound molecules. (Complex molecules such as those in dyes and DNA tend to be weakly bound.) Blue and violet dyes reflect those colors and, therefore, do not absorb these more energetic photons, thus suffering less molecular damage.



Why do the reds, yellows, and greens fade before the blues and violets when exposed to the Sun, as with this poster? The answer is related to photon energy. (credit: Deb Collins, Flickr)

Transparent materials, such as some glasses, do not absorb any visible light, because there is no energy step in the atoms or molecules that could absorb the light. Since individual photons interact with individual atoms, it is nearly impossible to have two photons absorbed simultaneously to reach a large energy step. Because of its lower photon energy, visible light can sometimes pass through many kilometers of a substance, while higher frequencies like UV, x ray, and γ rays are absorbed, because they have sufficient photon energy to ionize the material.

Example:
How Many Photons per Second Does a Typical Light Bulb Produce?
Assuming that 10.0% of a 100-W light bulb's energy output is in the visible range (typical for incandescent bulbs) with an average wavelength of 580 nm, calculate the number of visible photons emitted per second.
Strategy
Power is energy per unit time, and so if we can find the energy per photon, we can determine the number of photons per second. This will best be done in joules, since power is given in watts, which are joules per second.
Solution
The power in visible light production is 10.0% of 100 W, or 10.0 J/s. The energy of the average visible photon is found by substituting the given average wavelength into the formula
Equation:

$$E = \frac{\mathrm{hc}}{\lambda}.$$

This produces

Equation:

$$E = \frac{(6.63 \times 10^{-34}\ \mathrm{J \cdot s})(3.00 \times 10^{8}\ \mathrm{m/s})}{580 \times 10^{-9}\ \mathrm{m}} = 3.43 \times 10^{-19}\ \mathrm{J}.$$

The number of visible photons per second is thus

Equation:

$$\mathrm{photon/s} = \frac{10.0\ \mathrm{J/s}}{3.43 \times 10^{-19}\ \mathrm{J/photon}} = 2.92 \times 10^{19}\ \mathrm{photon/s}.$$

Discussion

This incredible number of photons per second is verification that individual photons are insignificant in ordinary human experience. It is also a verification of the correspondence principle—on the macroscopic scale, quantization becomes essentially continuous or classical. Finally, there are so many photons emitted by a 100-W lightbulb that it can be seen by the unaided eye many kilometers away.

Lower-Energy Photons

Infrared radiation (IR) has even lower photon energies than visible light and cannot significantly alter atoms and molecules. IR can be absorbed and emitted by atoms and molecules, particularly between closely spaced states. IR is extremely strongly absorbed by water, for example, because water molecules have many states separated by energies on the order of 10^{-5} eV to 10^{-2} eV, well within the IR and microwave energy ranges. This is why in the IR range, skin is almost jet black, with an emissivity near 1—there are many states in water molecules in the skin that can absorb a large range of IR photon energies. Not all molecules have this property. Air, for example, is nearly transparent to many IR frequencies.

Microwaves are the highest frequencies that can be produced by electronic circuits, although they are also produced naturally. Thus microwaves are similar to IR but do not extend to as high frequencies. There are states in water and other molecules that have the same frequency and energy as microwaves, typically about 10^{-5} eV. This is one reason why food absorbs microwaves more strongly than many other materials, making microwave ovens an efficient way of putting energy directly into food.

Photon energies for both IR and microwaves are so low that huge numbers of photons are involved in any significant energy transfer by IR or microwaves (such as warming yourself with a heat lamp or cooking pizza in the microwave). Visible light, IR, microwaves, and all lower frequencies cannot produce ionization with single photons and do not ordinarily have the hazards of higher frequencies. When visible, IR, or microwave radiation *is* hazardous, such as the inducement of cataracts by microwaves, the hazard is due to huge numbers of photons acting together (not to an accumulation of photons, such as sterilization by weak UV). The negative effects of visible, IR, or microwave radiation can be thermal effects, which could be produced by any heat source. But one difference is that at very high intensity, strong electric and magnetic fields can be produced by photons acting together. Such electromagnetic fields (EMF) can actually ionize materials.

Note:
Misconception Alert: High-Voltage Power Lines
Although some people think that living near high-voltage power lines is hazardous to one's health, ongoing studies of the transient field effects produced by these lines show their strengths to be insufficient to cause damage. Demographic studies also fail to show significant correlation of ill effects with high-voltage power lines. The American Physical Society issued a report over 10 years ago on power-line fields, which concluded that the scientific literature and reviews of panels show no consistent, significant link between cancer and power-line fields. They also felt that the "diversion of resources to eliminate a threat which has no persuasive scientific basis is disturbing."

It is virtually impossible to detect individual photons having frequencies below microwave frequencies, because of their low photon energy. But the photons are there. A continuous EM wave can be modeled as photons. At low frequencies, EM waves are generally treated as time- and position-varying electric and magnetic fields with no discernible quantization. This is another example of the correspondence principle in situations involving huge numbers of photons.

Note:
PhET Explorations: Color Vision
Make a whole rainbow by mixing red, green, and blue light. Change the wavelength of a monochromatic beam or filter white light. View the light as a solid beam, or see the individual photons.

https://phet.colorado.edu/sims/html/color-vision/latest/color-vision_en.html

Section Summary

- Photon energy is responsible for many characteristics of EM radiation, being particularly noticeable at high frequencies.
- Photons have both wave and particle characteristics.

Conceptual Questions

Exercise:

Problem: Why are UV, x rays, and γ rays called ionizing radiation?

Exercise:

Problem:

How can treating food with ionizing radiation help keep it from spoiling? UV is not very penetrating. What else could be used?

Exercise:

Problem:

Some television tubes are CRTs. They use an approximately 30-kV accelerating potential to send electrons to the screen, where the electrons stimulate phosphors to emit the light that forms the pictures we watch. Would you expect x rays also to be created?

Exercise:

Problem:

Tanning salons use “safe” UV with a longer wavelength than some of the UV in sunlight. This “safe” UV has enough photon energy to trigger the tanning mechanism. Is it likely to be able to cause cell damage and induce cancer with prolonged exposure?

Exercise:

Problem:

Your pupils dilate when visible light intensity is reduced. Does wearing sunglasses that lack UV blockers increase or decrease the UV hazard to your eyes? Explain.

Exercise:

Problem:

One could feel heat transfer in the form of infrared radiation from a large nuclear bomb detonated in the atmosphere 75 km from you. However, none of the profusely emitted x rays or γ rays reaches you. Explain.

Exercise:

Problem: Can a single microwave photon cause cell damage? Explain.

Exercise:

Problem:

In an x-ray tube, the maximum photon energy is given by $\text{hf} = \text{qV}$. Would it be technically more correct to say $\text{hf} = \text{qV} + \text{BE}$, where BE is the binding energy of electrons in the target anode? Why isn’t the energy stated the latter way?

Problems & Exercises

Exercise:

Problem:

What is the energy in joules and eV of a photon in a radio wave from an AM station that has a 1530-kHz broadcast frequency?

Solution:

6.34×10^{-9} eV, 1.01×10^{-27} J

Exercise:

Problem:

(a) Find the energy in joules and eV of photons in radio waves from an FM station that has a 90.0-MHz broadcast frequency. (b) What does this imply about the number of photons per second that the radio station must broadcast?

Exercise:

Problem: Calculate the frequency in hertz of a 1.00-MeV γ-ray photon.

Solution:

$2.42 \times 10^{20}\ \text{Hz}$

Exercise:

Problem:

(a) What is the wavelength of a 1.00-eV photon? (b) Find its frequency in hertz. (c) Identify the type of EM radiation.

Exercise:

Problem:

Do the unit conversions necessary to show that $\text{hc} = 1240\ \text{eV} \cdot \text{nm}$, as stated in the text.

Solution:
Equation:

$$\begin{aligned} \text{hc} &= \left(6.62607 \times 10^{-34}\ \text{J} \cdot \text{s}\right)\left(2.99792 \times 10^{8}\ \text{m/s}\right)\left(\frac{10^{9}\ \text{nm}}{1\ \text{m}}\right)\left(\frac{1.00000\ \text{eV}}{1.60218 \times 10^{-19}\ \text{J}}\right) \\ &= 1239.84\ \text{eV} \cdot \text{nm} \\ &\approx 1240\ \text{eV} \cdot \text{nm} \end{aligned}$$

Exercise:

Problem:

Confirm the statement in the text that the range of photon energies for visible light is 1.63 to 3.26 eV, given that the range of visible wavelengths is 380 to 760 nm.

Exercise:

Problem:

(a) Calculate the energy in eV of an IR photon of frequency $2.00 \times 10^{13}\ \text{Hz}$. (b) How many of these photons would need to be absorbed simultaneously by a tightly bound molecule to break it apart? (c) What is the energy in eV of a γ ray of frequency $3.00 \times 10^{20}\ \text{Hz}$? (d) How many tightly bound molecules could a single such γ ray break apart?

Solution:

(a) 0.0829 eV

(b) 121

(c) 1.24 MeV

(d) 1.24×10^5

Exercise:

Problem: Prove that, to three-digit accuracy, $h = 4.14 \times 10^{-15} \text{ eV} \cdot \text{s}$, as stated in the text.

Exercise:

Problem:

(a) What is the maximum energy in eV of photons produced in a CRT using a 25.0-kV accelerating potential, such as a color TV? (b) What is their frequency?

Solution:

(a) $25.0 \times 10^3 \text{ eV}$

(b) $6.04 \times 10^{18} \text{ Hz}$

Exercise:

Problem:

What is the accelerating voltage of an x-ray tube that produces x rays with a shortest wavelength of 0.0103 nm?

Exercise:

Problem:

(a) What is the ratio of power outputs by two microwave ovens having frequencies of 950 and 2560 MHz, if they emit the same number of photons per second? (b) What is the ratio of photons per second if they have the same power output?

Solution:

(a) 2.69

(b) 0.371

Exercise:

Problem:

How many photons per second are emitted by the antenna of a microwave oven, if its power output is 1.00 kW at a frequency of 2560 MHz?

Exercise:

Problem:

Some satellites use nuclear power. (a) If such a satellite emits a 1.00-W flux of γ rays having an average energy of 0.500 MeV, how many are emitted per second? (b) These γ rays affect other satellites. How far away must another satellite be to only receive one γ ray per second per square meter?

Solution:

(a) 1.25×10^{13} photons/s

(b) 997 km

Exercise:

Problem:

(a) If the power output of a 650-kHz radio station is 50.0 kW, how many photons per second are produced? (b) If the radio waves are broadcast uniformly in all directions, find the number of photons per second per square meter at a distance of 100 km. Assume no reflection from the ground or absorption by the air.

Exercise:

Problem:

How many x-ray photons per second are created by an x-ray tube that produces a flux of x rays having a power of 1.00 W? Assume the average energy per photon is 75.0 keV.

Solution:

8.33×10^{13} photons/s

Exercise:

Problem:

(a) How far away must you be from a 650-kHz radio station with power 50.0 kW for there to be only one photon per second per square meter? Assume no reflections or absorption, as if you were in deep outer space. (b) Discuss the implications for detecting intelligent life in other solar systems by detecting their radio broadcasts.

Exercise:

Problem:

Assuming that 10.0% of a 100-W light bulb’s energy output is in the visible range (typical for incandescent bulbs) with an average wavelength of 580 nm, and that the photons spread out uniformly and are not absorbed by the atmosphere, how far away would you be if 500 photons per second enter the 3.00-mm diameter pupil of your eye? (This number easily stimulates the retina.)

Solution:

181 km

Exercise:

Problem:Construct Your Own Problem

Consider a laser pen. Construct a problem in which you calculate the number of photons per second emitted by the pen. Among the things to be considered are the laser pen’s wavelength and power output. Your instructor may also wish for you to determine the minimum diffraction spreading in the beam and the number of photons per square centimeter the pen can project at some large distance. In this latter case, you will also need to consider the output size of the laser beam, the distance to the object being illuminated, and any absorption or scattering along the way.

Glossary

gamma ray
also γ-ray; highest-energy photon in the EM spectrum

ionizing radiation
radiation that ionizes materials that absorb it

x ray
EM photon between γ-ray and UV in energy

bremsstrahlung
German for *braking radiation*; produced when electrons are decelerated

characteristic x rays
x rays whose energy depends on the material they were produced in

ultraviolet radiation
UV; ionizing photons slightly more energetic than violet light

visible light
the range of photon energies the human eye can detect

infrared radiation
photons with energies slightly less than red light

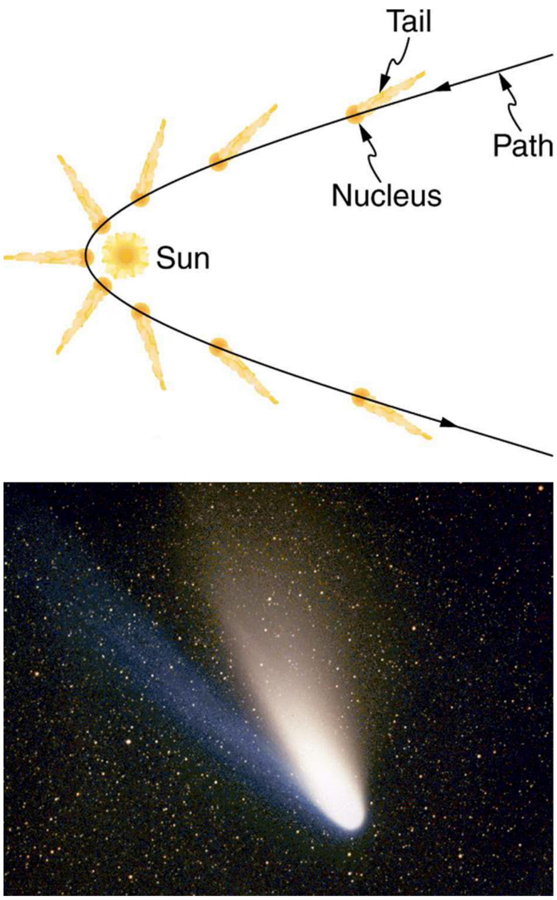
microwaves
photons with wavelengths on the order of a micron (μm)

Photon Momentum

- Relate the linear momentum of a photon to its energy or wavelength, and apply linear momentum conservation to simple processes involving the emission, absorption, or reflection of photons.
- Account qualitatively for the increase of photon wavelength that is observed, and explain the significance of the Compton wavelength.

Measuring Photon Momentum

The quantum of EM radiation we call a **photon** has properties analogous to those of particles we can see, such as grains of sand. A photon interacts as a unit in collisions or when absorbed, rather than as an extensive wave. Massive quanta, like electrons, also act like macroscopic particles—something we expect, because they are the smallest units of matter. Particles carry momentum as well as energy. Despite photons having no mass, there has long been evidence that EM radiation carries momentum. (Maxwell and others who studied EM waves predicted that they would carry momentum.) It is now a well-established fact that photons *do* have momentum. In fact, photon momentum is suggested by the photoelectric effect, where photons knock electrons out of a substance. [link] shows macroscopic evidence of photon momentum.



The tails of the Hale-Bopp comet point away from the Sun, evidence that light has momentum. Dust emanating from the body of the comet forms this tail. Particles of dust are pushed away from the Sun by light reflecting from them. The blue ionized gas tail is also produced by photons interacting with atoms in the comet material. (credit: Geoff Chester, U.S. Navy, via Wikimedia Commons)

[link] shows a comet with two prominent tails. What most people do not know about the tails is that they always point *away* from the Sun rather than trailing behind the comet (like the tail of Bo Peep’s sheep). Comet tails are composed of gases and dust evaporated from the body of the comet and ionized gas. The dust particles recoil away from the Sun when photons scatter from them. Evidently, photons carry momentum in the direction of their motion (away from the Sun), and some of this momentum is transferred to dust particles in collisions. Gas atoms and molecules in the blue tail are most affected by other particles of radiation, such as protons and electrons emanating from the Sun, rather than by the momentum of photons.

Note:

Connections: Conservation of Momentum

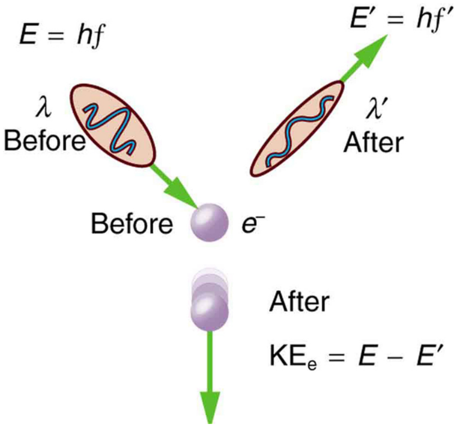
Not only is momentum conserved in all realms of physics, but all types of particles are found to have momentum. We expect particles with mass to have momentum, but now we see that massless particles including photons also carry momentum.

Momentum is conserved in quantum mechanics just as it is in relativity and classical physics. Some of the earliest direct experimental evidence of this came from scattering of x-ray photons by electrons in substances, named Compton scattering after the American physicist Arthur H. Compton (1892–1962). Around 1923, Compton observed that x rays scattered from materials had a decreased energy and correctly analyzed this as being due to the scattering of photons from electrons. This phenomenon could be handled as a collision between two particles—a photon and an electron at rest in the material. Energy and momentum are conserved in the collision. (See [link]) He won a Nobel Prize in 1929 for the discovery of this scattering, now called the **Compton effect**, because it helped prove that **photon momentum** is given by

Equation:

$$p = \frac{h}{\lambda},$$

where h is Planck's constant and λ is the photon wavelength. (Note that relativistic momentum given as $p = \gamma \mathrm{mu}$ is valid only for particles having mass.)



The Compton effect is the name given to the scattering of a photon by an electron. Energy and momentum are conserved, resulting in a reduction of both for the scattered photon. Studying this effect, Compton verified that photons have momentum.

We can see that photon momentum is small, since $p = h/\lambda$ and h is very small. It is for this reason that we do not ordinarily observe photon momentum. Our mirrors do not recoil when light reflects from them (except perhaps in cartoons). Compton saw the effects of photon momentum because he was observing x rays, which have a small wavelength and a relatively large momentum, interacting with the lightest of particles, the electron.

Example:
Electron and Photon Momentum Compared

(a) Calculate the momentum of a visible photon that has a wavelength of 500 nm. (b) Find the velocity of an electron having the same momentum. (c) What is the energy of the electron, and how does it compare with the energy of the photon?

Strategy

Finding the photon momentum is a straightforward application of its definition: $p=\frac{h}{\lambda}$. If we find the photon momentum is small, then we can assume that an electron with the same momentum will be nonrelativistic, making it easy to find its velocity and kinetic energy from the classical formulas.

Solution for (a)

Photon momentum is given by the equation:

Equation:

$$p=\frac{h}{\lambda}.$$

Entering the given photon wavelength yields

Equation:

$$p=\frac{6.63\times10^{-34}\text{ J}\cdot\text{s}}{500\times10^{-9}\text{ m}}=1.33\times10^{-27}\text{ kg}\cdot\text{m/s}.$$

Solution for (b)

Since this momentum is indeed small, we will use the classical expression $p=\text{mv}$ to find the velocity of an electron with this momentum. Solving for v and using the known value for the mass of an electron gives

Equation:

$$v=\frac{p}{m}=\frac{1.33\times 10^{-27}\ \mathrm{kg}\cdot\mathrm{m/s}}{9.11\times 10^{-31}\ \mathrm{kg}}=1460\ \mathrm{m/s}\approx 1460\ \mathrm{m/s}.$$

Solution for (c)
The electron has kinetic energy, which is classically given by
Equation:

$$\mathrm{KE}_e=\frac{1}{2}mv^2.$$

Thus,
Equation:

$$\mathrm{KE}_e=\frac{1}{2}(9.11\times 10^{-3}\ \mathrm{kg})(1455\ \mathrm{m/s})^2=9.64\times 10^{-25}\ \mathrm{J}.$$

Converting this to eV by multiplying by $(1\ \mathrm{eV})/(1.602\times 10^{-19}\ \mathrm{J})$ yields
Equation:

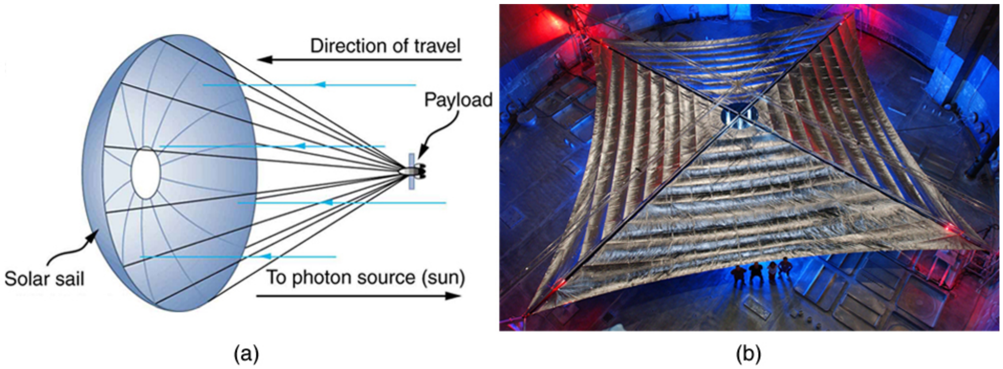
$$\mathrm{KE}_e=6.02\times 10^{-6}\ \mathrm{eV}.$$

The photon energy E is
Equation:

$$E=\frac{\mathrm{hc}}{\lambda}=\frac{1240\ \mathrm{eV}\cdot\mathrm{nm}}{500\ \mathrm{nm}}=2.48\ \mathrm{eV},$$

which is about five orders of magnitude greater.
Discussion
Photon momentum is indeed small. Even if we have huge numbers of them, the total momentum they carry is small. An electron with the same momentum has a 1460 m/s velocity, which is clearly nonrelativistic. A more massive particle with the same momentum would have an even smaller velocity. This is borne out by the fact that it takes far less energy to give an electron the same momentum as a photon. But on a quantum-mechanical scale, especially for high-energy photons interacting with small

masses, photon momentum is significant. Even on a large scale, photon momentum can have an effect if there are enough of them and if there is nothing to prevent the slow recoil of matter. Comet tails are one example, but there are also proposals to build space sails that use huge low-mass mirrors (made of aluminized Mylar) to reflect sunlight. In the vacuum of space, the mirrors would gradually recoil and could actually take spacecraft from place to place in the solar system. (See [link].)



(a) Space sails have been proposed that use the momentum of sunlight reflecting from gigantic low-mass sails to propel spacecraft about the solar system. A Russian test model of this (the Cosmos 1) was launched in 2005, but did not make it into orbit due to a rocket failure. (b) A U.S. version of this, labeled LightSail-1, is scheduled for trial launches in the first part of this decade. It will have a 40-m^2 sail. (credit: Kim Newton/NASA)

Relativistic Photon Momentum

There is a relationship between photon momentum p and photon energy E that is consistent with the relation given previously for the relativistic total energy of a particle as $E^2 = (\mathrm{pc})^2 + (\mathrm{mc})^2$. We know m is zero for a photon, but p is not, so that $E^2 = (\mathrm{pc})^2 + (\mathrm{mc})^2$ becomes

Equation:

$$E=\text{pc},$$

or

Equation:

$$p=\frac{E}{c}\text{ (photons)}.$$

To check the validity of this relation, note that $E=\text{hc}/\lambda$ for a photon. Substituting this into $p=E/c$ yields

Equation:

$$p=(\text{hc}/\lambda)/c=\frac{h}{\lambda},$$

as determined experimentally and discussed above. Thus, $p=E/c$ is equivalent to Compton’s result $p=h/\lambda$. For a further verification of the relationship between photon energy and momentum, see [link].

Note:

Photon Detectors

Almost all detection systems talked about thus far—eyes, photographic plates, photomultiplier tubes in microscopes, and CCD cameras—rely on particle-like properties of photons interacting with a sensitive area. A change is caused and either the change is cascaded or zillions of points are recorded to form an image we detect. These detectors are used in biomedical imaging systems, and there is ongoing research into improving the efficiency of receiving photons, particularly by cooling detection systems and reducing thermal effects.

Example:
Photon Energy and Momentum

Show that $p = E/c$ for the photon considered in the [link].

Strategy

We will take the energy E found in [link], divide it by the speed of light, and see if the same momentum is obtained as before.

Solution

Given that the energy of the photon is 2.48 eV and converting this to joules, we get

Equation:

$$p = \frac{E}{c} = \frac{(2.48 \text{ eV})(1.60 \times 10^{-19} \text{ J/eV})}{3.00 \times 10^{8} \text{ m/s}} = 1.33 \times 10^{-27} \text{ kg} \cdot \text{m/s}.$$

Discussion

This value for momentum is the same as found before (note that unrounded values are used in all calculations to avoid even small rounding errors), an expected verification of the relationship $p = E/c$. This also means the relationship between energy, momentum, and mass given by $E^2 = (\text{pc})^2 + (\text{mc})^2$ applies to both matter and photons. Once again, note that p is not zero, even when m is.

Note:
Problem-Solving Suggestion

Note that the forms of the constants $h = 4.14 \times 10^{-15} \text{ eV} \cdot \text{s}$ and $\text{hc} = 1240 \text{ eV} \cdot \text{nm}$ may be particularly useful for this section’s Problems and Exercises.

Section Summary

- Photons have momentum, given by $p = \frac{h}{\lambda}$, where λ is the photon wavelength.

- Photon energy and momentum are related by $p = \frac{E}{c}$, where $E = \text{hf} = \text{hc}/\lambda$ for a photon.

Conceptual Questions

Exercise:

Problem:

Which formula may be used for the momentum of all particles, with or without mass?

Exercise:

Problem:

Is there any measurable difference between the momentum of a photon and the momentum of matter?

Exercise:

Problem:

Why don’t we feel the momentum of sunlight when we are on the beach?

Problems & Exercises

Exercise:

Problem:

(a) Find the momentum of a 4.00-cm-wavelength microwave photon.
(b) Discuss why you expect the answer to (a) to be very small.

Solution:

(a) $1.66 \times 10^{-32}\ \text{kg} \cdot \text{m/s}$

(b) The wavelength of microwave photons is large, so the momentum they carry is very small.

Exercise:

Problem:

(a) What is the momentum of a 0.0100-nm-wavelength photon that could detect details of an atom? (b) What is its energy in MeV?

Exercise:

Problem:

(a) What is the wavelength of a photon that has a momentum of $5.00 \times 10^{-29}\ \text{kg} \cdot \text{m/s}$? (b) Find its energy in eV.

Solution:

(a) 13.3 µm

(b) 9.38×10^{-2} eV

Exercise:

Problem:

(a) A γ-ray photon has a momentum of $8.00 \times 10^{-21}\ \text{kg} \cdot \text{m/s}$. What is its wavelength? (b) Calculate its energy in MeV.

Exercise:

Problem:

(a) Calculate the momentum of a photon having a wavelength of $2.50\ \mu\text{m}$. (b) Find the velocity of an electron having the same momentum. (c) What is the kinetic energy of the electron, and how does it compare with that of the photon?

Solution:

(a) $2.65 \times 10^{-28}\ \text{kg} \cdot \text{m/s}$

(b) 291 m/s

(c) electron 3.86×10^{-26} J, photon 7.96×10^{-20} J, ratio 2.06×10^{6}

Exercise:

Problem:

Repeat the previous problem for a 10.0-nm-wavelength photon.

Exercise:

Problem:

(a) Calculate the wavelength of a photon that has the same momentum as a proton moving at 1.00% of the speed of light. (b) What is the energy of the photon in MeV? (c) What is the kinetic energy of the proton in MeV?

Solution:

(a) 1.32×10^{-13} m

(b) 9.39 MeV

(c) 4.70×10^{-2} MeV

Exercise:

Problem:

(a) Find the momentum of a 100-keV x-ray photon. (b) Find the equivalent velocity of a neutron with the same momentum. (c) What is the neutron’s kinetic energy in keV?

Exercise:

Problem:

Take the ratio of relativistic rest energy, $E = \gamma mc^2$, to relativistic momentum, $p = \gamma \text{mu}$, and show that in the limit that mass approaches zero, you find $E/p = c$.

Solution:

$E = \gamma mc^2$ and $P = \gamma mu$, so

Equation:

$$\frac{E}{P} = \frac{\gamma mc^2}{\gamma mu} = \frac{c^2}{u}.$$

As the mass of particle approaches zero, its velocity u will approach c, so that the ratio of energy to momentum in this limit is

Equation:

$$\lim_{m \to 0} \frac{E}{P} = \frac{c^2}{c} = c$$

which is consistent with the equation for photon energy.

Exercise:

Problem: Construct Your Own Problem

Consider a space sail such as mentioned in [link]. Construct a problem in which you calculate the light pressure on the sail in $\mathrm{N/m^2}$ produced by reflecting sunlight. Also calculate the force that could be produced and how much effect that would have on a spacecraft. Among the things to be considered are the intensity of sunlight, its average wavelength, the number of photons per square meter this implies, the area of the space sail, and the mass of the system being accelerated.

Exercise:

Problem: Unreasonable Results

A car feels a small force due to the light it sends out from its headlights, equal to the momentum of the light divided by the time in which it is emitted. (a) Calculate the power of each headlight, if they exert a total force of 2.00×10^{-2} N backward on the car. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $3.00 \times 10^6 \text{ W}$

(b) Headlights are way too bright.

(c) Force is too large.

Glossary

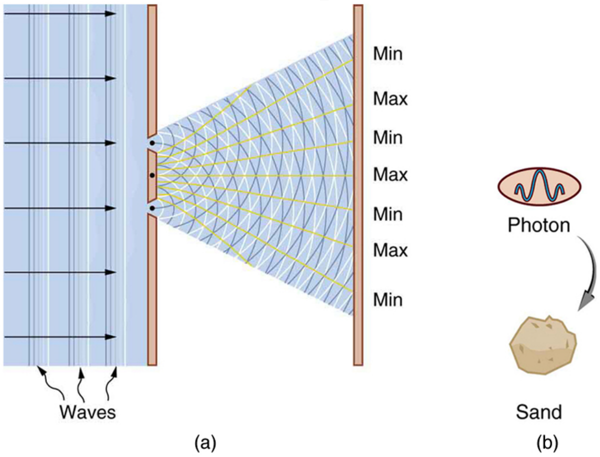
photon momentum
: the amount of momentum a photon has, calculated by $p = \frac{h}{\lambda} = \frac{E}{c}$

Compton effect
: the phenomenon whereby x rays scattered from materials have decreased energy

The Particle-Wave Duality

- Explain what the term particle-wave duality means, and why it is applied to EM radiation.

We have long known that EM radiation is a wave, capable of interference and diffraction. We now see that light can be modeled as photons, which are massless particles. This may seem contradictory, since we ordinarily deal with large objects that never act like both wave and particle. An ocean wave, for example, looks nothing like a rock. To understand small-scale phenomena, we make analogies with the large-scale phenomena we observe directly. When we say something behaves like a wave, we mean it shows interference effects analogous to those seen in overlapping water waves. (See [link].) Two examples of waves are sound and EM radiation. When we say something behaves like a particle, we mean that it interacts as a discrete unit with no interference effects. Examples of particles include electrons, atoms, and photons of EM radiation. How do we talk about a phenomenon that acts like both a particle and a wave?



(a) The interference pattern for light through a double slit is a wave property understood by analogy to water waves. (b) The properties of photons having quantized energy and momentum and acting as a concentrated unit are understood by analogy to macroscopic particles.

There is no doubt that EM radiation interferes and has the properties of wavelength and frequency. There is also no doubt that it behaves as particles—photons with discrete energy. We call this twofold nature the **particle-wave duality**, meaning that EM radiation has both particle and wave properties. This so-called duality is simply a term for properties of the photon analogous to phenomena we can observe directly, on a macroscopic scale. If this term seems strange, it is because we do not ordinarily observe details on the quantum level directly, and our observations yield either particle *or* wavelike properties, but never both simultaneously.

Since we have a particle-wave duality for photons, and since we have seen connections between photons and matter in that both have momentum, it is reasonable to ask whether there is a particle-wave duality for matter as well. If the EM radiation we once thought to be a pure wave has particle properties, is it possible that matter has wave properties? The answer is yes. The consequences are tremendous, as we will begin to see in the next section.

Note:
PhET Explorations: Quantum Wave Interference
When do photons, electrons, and atoms behave like particles and when do they behave like waves? Watch waves spread out and interfere as they pass through a double slit, then get detected on a screen as tiny dots. Use quantum detectors to explore how measurements change the waves and the patterns they produce on the screen.

Quantum Wave Interference

Section Summary

- EM radiation can behave like either a particle or a wave.
- This is termed particle-wave duality.

Glossary

particle-wave duality
the property of behaving like either a particle or a wave; the term for the phenomenon that all particles have wave characteristics

The Wave Nature of Matter

- Describe the Davisson-Germer experiment, and explain how it provides evidence for the wave nature of electrons.

De Broglie Wavelength

In 1923 a French physics graduate student named Prince Louis-Victor de Broglie (1892–1987) made a radical proposal based on the hope that nature is symmetric. If EM radiation has both particle and wave properties, then nature would be symmetric if matter also had both particle and wave properties. If what we once thought of as an unequivocal wave (EM radiation) is also a particle, then what we think of as an unequivocal particle (matter) may also be a wave. De Broglie's suggestion, made as part of his doctoral thesis, was so radical that it was greeted with some skepticism. A copy of his thesis was sent to Einstein, who said it was not only probably correct, but that it might be of fundamental importance. With the support of Einstein and a few other prominent physicists, de Broglie was awarded his doctorate.

De Broglie took both relativity and quantum mechanics into account to develop the proposal that *all particles have a wavelength*, given by
Equation:

$$\lambda = \frac{h}{p} \text{ (matter and photons)},$$

where h is Planck's constant and p is momentum. This is defined to be the **de Broglie wavelength**. (Note that we already have this for photons, from the equation $p = h/\lambda$.) The hallmark of a wave is interference. If matter is a wave, then it must exhibit constructive and destructive interference. Why isn't this ordinarily observed? The answer is that in order to see significant interference effects, a wave must interact with an object about the same size as its wavelength. Since h is very small, λ is also small, especially for macroscopic objects. A 3-kg bowling ball moving at 10 m/s, for example, has

Equation:

$$\lambda = h/p = (6.63 \times 10^{-34}\ \text{J·s})/[(3\ \text{kg})(10\ \text{m/s}\ )]= 2 \times 10^{-35}\ \text{m}.$$

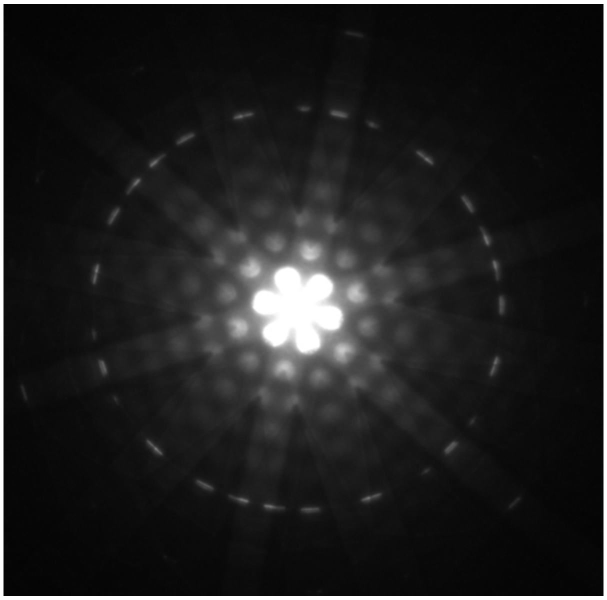
This means that to see its wave characteristics, the bowling ball would have to interact with something about 10^{-35} m in size—far smaller than anything known. When waves interact with objects much larger than their wavelength, they show negligible interference effects and move in straight lines (such as light rays in geometric optics). To get easily observed interference effects from particles of matter, the longest wavelength and hence smallest mass possible would be useful. Therefore, this effect was first observed with electrons.

American physicists Clinton J. Davisson and Lester H. Germer in 1925 and, independently, British physicist G. P. Thomson (son of J. J. Thomson, discoverer of the electron) in 1926 scattered electrons from crystals and found diffraction patterns. These patterns are exactly consistent with interference of electrons having the de Broglie wavelength and are somewhat analogous to light interacting with a diffraction grating. (See [link].)

Note:
Connections: Waves
All microscopic particles, whether massless, like photons, or having mass, like electrons, have wave properties. The relationship between momentum and wavelength is fundamental for all particles.

De Broglie's proposal of a wave nature for all particles initiated a remarkably productive era in which the foundations for quantum mechanics were laid. In 1926, the Austrian physicist Erwin Schrödinger (1887–1961) published four papers in which the wave nature of particles was treated explicitly with wave equations. At the same time, many others began important work. Among them was German physicist Werner Heisenberg

(1901–1976) who, among many other contributions to quantum mechanics, formulated a mathematical treatment of the wave nature of matter that used matrices rather than wave equations. We will deal with some specifics in later sections, but it is worth noting that de Broglie's work was a watershed for the development of quantum mechanics. De Broglie was awarded the Nobel Prize in 1929 for his vision, as were Davisson and G. P. Thomson in 1937 for their experimental verification of de Broglie's hypothesis.



This diffraction pattern was obtained for electrons diffracted by crystalline silicon. Bright regions are those of constructive interference, while dark regions are those of destructive interference. (credit: Ndthe, Wikimedia Commons)

Example:

Electron Wavelength versus Velocity and Energy

For an electron having a de Broglie wavelength of 0.167 nm (appropriate for interacting with crystal lattice structures that are about this size): (a) Calculate the electron's velocity, assuming it is nonrelativistic. (b) Calculate the electron's kinetic energy in eV.

Strategy

For part (a), since the de Broglie wavelength is given, the electron's velocity can be obtained from $\lambda = h/p$ by using the nonrelativistic formula for momentum, $p = \text{mv}$. For part (b), once v is obtained (and it has been verified that v is nonrelativistic), the classical kinetic energy is simply $(1/2)mv^2$.

Solution for (a)

Substituting the nonrelativistic formula for momentum ($p = \text{mv}$) into the de Broglie wavelength gives

Equation:

$$\lambda = \frac{h}{p} = \frac{h}{\text{mv}}.$$

Solving for v gives

Equation:

$$v = \frac{h}{m\lambda}.$$

Substituting known values yields

Equation:

$$v = \frac{6.63 \times 10^{-34}\ \text{J} \cdot \text{s}}{(9.11 \times 10^{-31}\ \text{kg})(0.167 \times 10^{-9}\ \text{m})} = 4.36 \times 10^{6}\ \text{m/s}.$$

Solution for (b)

While fast compared with a car, this electron's speed is not highly relativistic, and so we can comfortably use the classical formula to find the electron's kinetic energy and convert it to eV as requested.

Equation:

$$\begin{aligned}\text{KE} &= \frac{1}{2}mv^2 \\ &= \frac{1}{2}(9.11 \times 10^{-31}\ \text{kg})(4.36 \times 10^{6}\ \text{m/s})^2 \\ &= (86.4 \times 10^{-18}\ \text{J})\left(\frac{1\ \text{eV}}{1.602 \times 10^{-19}\ \text{J}}\right) \\ &= 54.0\ \text{eV}\end{aligned}$$

Discussion

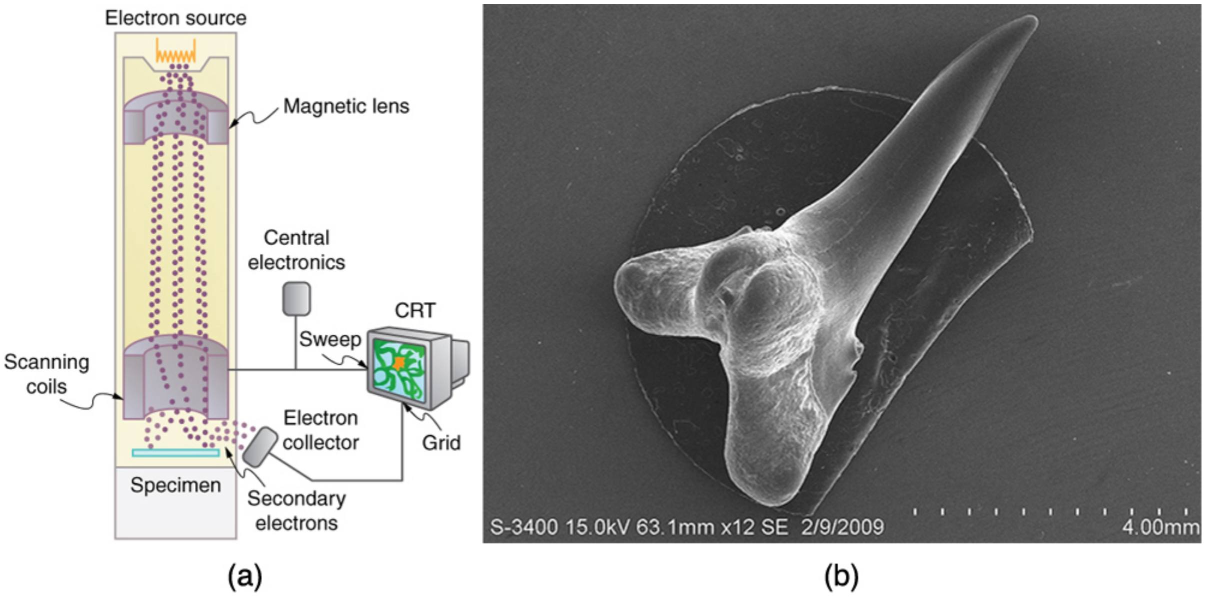
This low energy means that these 0.167-nm electrons could be obtained by accelerating them through a 54.0-V electrostatic potential, an easy task. The results also confirm the assumption that the electrons are nonrelativistic, since their velocity is just over 1% of the speed of light and the kinetic energy is about 0.01% of the rest energy of an electron (0.511 MeV). If the electrons had turned out to be relativistic, we would have had to use more involved calculations employing relativistic formulas.

Electron Microscopes

One consequence or use of the wave nature of matter is found in the electron microscope. As we have discussed, there is a limit to the detail observed with any probe having a wavelength. Resolution, or observable detail, is limited to about one wavelength. Since a potential of only 54 V can produce electrons with sub-nanometer wavelengths, it is easy to get electrons with much smaller wavelengths than those of visible light (hundreds of nanometers). Electron microscopes can, thus, be constructed to detect much smaller details than optical microscopes. (See [link].)

There are basically two types of electron microscopes. The transmission electron microscope (TEM) accelerates electrons that are emitted from a hot filament (the cathode). The beam is broadened and then passes through the sample. A magnetic lens focuses the beam image onto a fluorescent screen, a photographic plate, or (most probably) a CCD (light sensitive camera), from which it is transferred to a computer. The TEM is similar to the optical microscope, but it requires a thin sample examined in a vacuum. However it can resolve details as small as 0.1 nm (10^{-10} m), providing magnifications of 100 million times the size of the original object. The TEM has allowed us to see individual atoms and structure of cell nuclei.

The scanning electron microscope (SEM) provides images by using secondary electrons produced by the primary beam interacting with the surface of the sample (see [link]). The SEM also uses magnetic lenses to focus the beam onto the sample. However, it moves the beam around electrically to "scan" the sample in the *x* and *y* directions. A CCD detector is used to process the data for each electron position, producing images like the one at the beginning of this chapter. The SEM has the advantage of not requiring a thin sample and of providing a 3-D view. However, its resolution is about ten times less than a TEM.



Schematic of a scanning electron microscope (SEM) (a) used to observe small details, such as those seen in this image of a tooth of a *Himipristis*, a type of shark (b). (credit: Dallas Krentzel, Flickr)

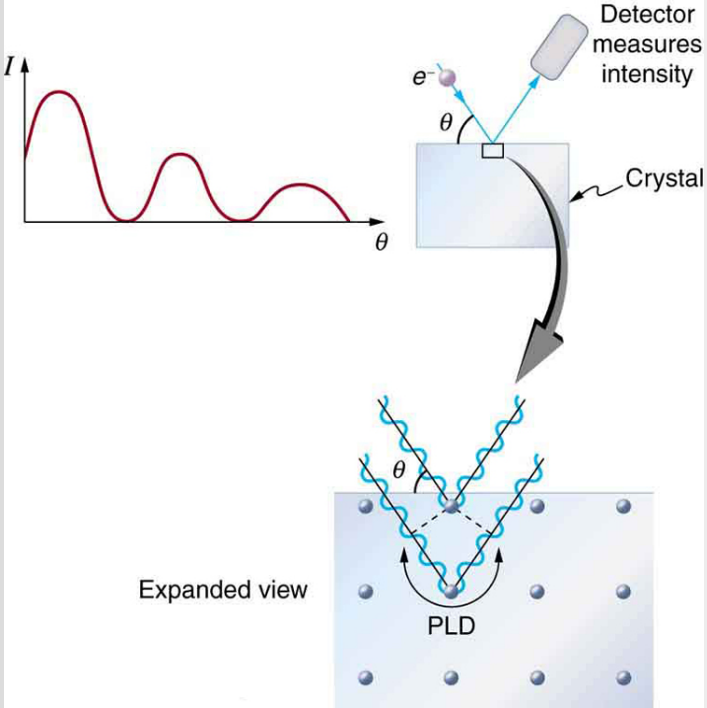
Electrons were the first particles with mass to be directly confirmed to have the wavelength proposed by de Broglie. Subsequently, protons, helium nuclei, neutrons, and many others have been observed to exhibit interference when they interact with objects having sizes similar to their de Broglie wavelength. The de Broglie wavelength for massless particles was well established in the 1920s for photons, and it has since been observed that all massless particles have a de Broglie wavelength $\lambda = h/p$ The wave nature of all particles is a universal characteristic of nature. We shall see in following sections that implications of the de Broglie wavelength include the quantization of energy in atoms and molecules, and an alteration of our basic view of nature on the microscopic scale. The next section, for example, shows that there are limits to the precision with which we may make predictions, regardless of how hard we try. There are even limits to the precision with which we may measure an object's location or energy.

Note:

Making Connections: A Submicroscopic Diffraction Grating

The wave nature of matter allows it to exhibit all the characteristics of other, more familiar, waves. Diffraction gratings, for example, produce diffraction patterns for light that depend on grating spacing and the wavelength of the light. This effect, as with most wave phenomena, is most pronounced when the wave interacts with objects having a size similar to its wavelength. For gratings, this is the spacing between multiple slits.) When electrons interact with a system having a spacing similar to the electron wavelength, they show the same types of interference patterns as light does for diffraction gratings, as shown at top left in [link].

Atoms are spaced at regular intervals in a crystal as parallel planes, as shown in the bottom part of [link]. The spacings between these planes act like the openings in a diffraction grating. At certain incident angles, the paths of electrons scattering from successive planes differ by one wavelength and, thus, interfere constructively. At other angles, the path length differences are not an integral wavelength, and there is partial to total destructive interference. This type of scattering from a large crystal with well-defined lattice planes can produce dramatic interference patterns. It is called *Bragg reflection*, for the father-and-son team who first explored and analyzed it in some detail. The expanded view also shows the path-length differences and indicates how these depend on incident angle θ in a manner similar to the diffraction patterns for x rays reflecting from a crystal.



The diffraction pattern at top left is produced by scattering electrons from a crystal and is graphed as a function of incident angle relative to the regular array of atoms in a crystal, as shown at bottom. Electrons scattering from the second layer of atoms travel farther than those scattered from the top layer. If the path length difference (PLD) is an integral wavelength, there is constructive interference.

Let us take the spacing between parallel planes of atoms in the crystal to be d. As mentioned, if the path length difference (PLD) for the electrons is a whole number of wavelengths, there will be constructive interference—that is, $\text{PLD} = n\lambda (n = 1, 2, 3, \ldots)$. Because $\text{AB} = \text{BC} = d \sin \theta$, we have constructive interference when $n\lambda = 2d$ sin θ. This relationship is

called the *Bragg equation* and applies not only to electrons but also to x rays.
The wavelength of matter is a submicroscopic characteristic that explains a macroscopic phenomenon such as Bragg reflection. Similarly, the wavelength of light is a submicroscopic characteristic that explains the macroscopic phenomenon of diffraction patterns.

Section Summary

- Particles of matter also have a wavelength, called the de Broglie wavelength, given by $\lambda = \frac{h}{p}$, where p is momentum.
- Matter is found to have the same *interference characteristics* as any other wave.

Conceptual Questions

Exercise:

Problem:

How does the interference of water waves differ from the interference of electrons? How are they analogous?

Exercise:

Problem: Describe one type of evidence for the wave nature of matter.

Exercise:

Problem:

Describe one type of evidence for the particle nature of EM radiation.

Problems & Exercises

Exercise:

Problem:

At what velocity will an electron have a wavelength of 1.00 m?

Solution:

$7.28 \times 10^{-4}\ \text{m}$

Exercise:

Problem:

What is the wavelength of an electron moving at 3.00% of the speed of light?

Exercise:

Problem:

At what velocity does a proton have a 6.00-fm wavelength (about the size of a nucleus)? Assume the proton is nonrelativistic. (1 femtometer = 10^{-15} m.)

Solution:

$6.62 \times 10^{7}\ \text{m/s}$

Exercise:

Problem:

What is the velocity of a 0.400-kg billiard ball if its wavelength is 7.50 cm (large enough for it to interfere with other billiard balls)?

Exercise:

Problem:

Find the wavelength of a proton moving at 1.00% of the speed of light.

Solution:

$1.32 \times 10^{-13}\ \text{m}$

Exercise:

Problem:

Experiments are performed with ultracold neutrons having velocities as small as 1.00 m/s. (a) What is the wavelength of such a neutron? (b) What is its kinetic energy in eV?

Exercise:

Problem:

(a) Find the velocity of a neutron that has a 6.00-fm wavelength (about the size of a nucleus). Assume the neutron is nonrelativistic. (b) What is the neutron's kinetic energy in MeV?

Solution:

(a) $6.62 \times 10^{7}\ \text{m/s}$

(b) $22.9\ \text{MeV}$

Exercise:

Problem:

What is the wavelength of an electron accelerated through a 30.0-kV potential, as in a TV tube?

Exercise:

Problem:

What is the kinetic energy of an electron in a TEM having a 0.0100-nm wavelength?

Solution:
Equation:15.1 keV

Exercise:

Problem:

(a) Calculate the velocity of an electron that has a wavelength of $1.00\ \mu\text{m}$. (b) Through what voltage must the electron be accelerated to have this velocity?

Exercise:

Problem:

The velocity of a proton emerging from a Van de Graaff accelerator is 25.0% of the speed of light. (a) What is the proton’s wavelength? (b) What is its kinetic energy, assuming it is nonrelativistic? (c) What was the equivalent voltage through which it was accelerated?

Solution:

(a) 5.29 fm

(b) $4.70 \times 10^{-12}\ \text{J}$

(c) 29.4 MV

Exercise:

Problem:

The kinetic energy of an electron accelerated in an x-ray tube is 100 keV. Assuming it is nonrelativistic, what is its wavelength?

Exercise:

Problem: Unreasonable Results

(a) Assuming it is nonrelativistic, calculate the velocity of an electron with a 0.100-fm wavelength (small enough to detect details of a nucleus). (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $7.28 \times 10^{12}\ \mathrm{m/s}$

(b) This is thousands of times the speed of light (an impossibility).

(c) The assumption that the electron is non-relativistic is unreasonable at this wavelength.

Glossary

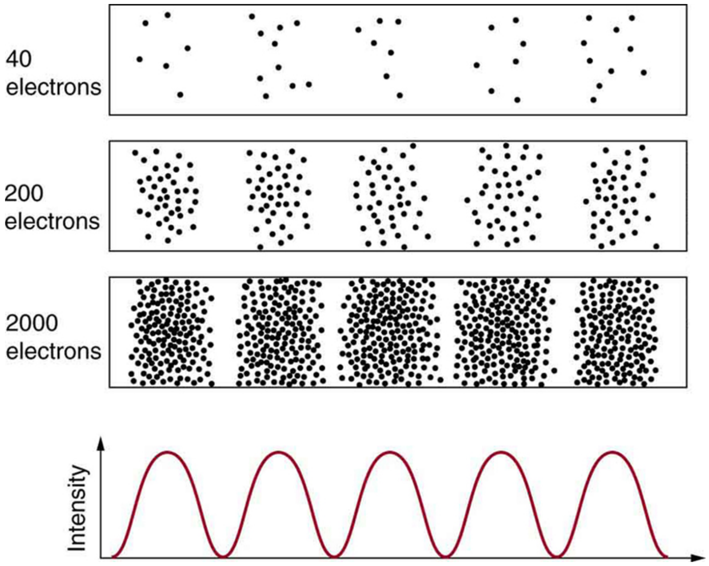
de Broglie wavelength
: the wavelength possessed by a particle of matter, calculated by $\lambda = h/p$

Probability: The Heisenberg Uncertainty Principle

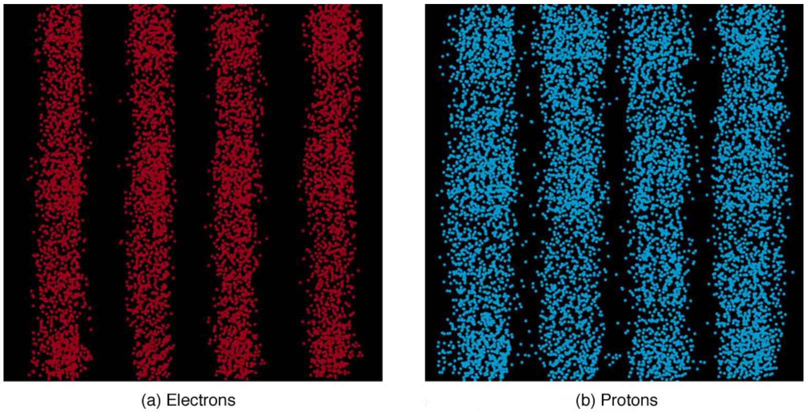
- Use both versions of Heisenberg’s uncertainty principle in calculations.
- Explain the implications of Heisenberg’s uncertainty principle for measurements.

Probability Distribution

Matter and photons are waves, implying they are spread out over some distance. What is the position of a particle, such as an electron? Is it at the center of the wave? The answer lies in how you measure the position of an electron. Experiments show that you will find the electron at some definite location, unlike a wave. But if you set up exactly the same situation and measure it again, you will find the electron in a different location, often far outside any experimental uncertainty in your measurement. Repeated measurements will display a statistical distribution of locations that appears wavelike. (See [link].)



The building up of the diffraction pattern of electrons scattered from a crystal surface. Each electron arrives at a definite location, which cannot be precisely predicted. The overall distribution shown at the bottom can be predicted as the diffraction of waves having the de Broglie wavelength of the electrons.



Double-slit interference for electrons (a) and protons (b) is identical for equal wavelengths and equal slit separations. Both patterns are probability distributions in the sense that they are built up by individual particles traversing the apparatus, the paths of which are not individually predictable.

After de Broglie proposed the wave nature of matter, many physicists, including Schrödinger and Heisenberg, explored the consequences. The idea quickly emerged that, *because of its wave character, a particle's trajectory and destination cannot be precisely predicted for each particle individually*. However, each particle goes to a definite place (as illustrated in [link]). After compiling enough data, you get a distribution related to the particle’s wavelength and diffraction pattern. There is a certain *probability* of finding the particle at a given location, and the overall pattern is called a **probability distribution**. Those who developed quantum mechanics devised equations that predicted the probability distribution in various circumstances.

It is somewhat disquieting to think that you cannot predict exactly where an individual particle will go, or even follow it to its destination. Let us explore what happens if we try to follow a particle. Consider the double-slit patterns obtained for electrons and photons in [link]. First, we note that these patterns are identical, following $d \sin \theta = m\lambda$, the equation for double-slit constructive interference developed in Photon Energies and the Electromagnetic Spectrum, where d is the slit separation and λ is the electron or photon wavelength.

Both patterns build up statistically as individual particles fall on the detector. This can be observed for photons or electrons—for now, let us concentrate on electrons. You might imagine that the electrons are interfering with one another as any waves do. To test this, you can lower the intensity until there is never more than one electron between the slits and the screen. The same interference pattern builds up! This implies that a particle’s probability distribution spans both slits, and the particles actually interfere with themselves. Does this also mean that the electron goes through both slits? An electron is a basic unit of matter that is not divisible. But it is a fair question, and so we should look to see if the electron traverses one slit or the other, or both. One possibility is to have coils around the slits that detect charges moving through them. What is observed is that an electron always goes through one slit or the other; it does not split to go through both. But there is a catch. If you determine that the electron went through one of the slits, you no longer get a double slit pattern—instead, you get single slit interference. There is no escape by using another method of determining which slit the electron went through. Knowing the particle went through one slit forces a single-slit pattern. If you do not observe which slit the electron goes through, you obtain a double-slit pattern.

Heisenberg Uncertainty

How does knowing which slit the electron passed through change the pattern? The answer is fundamentally important—*measurement affects the system being observed*. Information can be lost, and in some cases it is impossible to measure two physical quantities simultaneously to exact precision. For example, you can measure the position of a moving electron by scattering light or other electrons from it. Those probes have momentum themselves, and by scattering from the electron, they change its momentum *in a manner that loses information*. There is a limit to absolute knowledge, even in principle.

Werner Heisenberg was one of the best of those physicists who developed early quantum mechanics. Not only did his work enable a description of nature on the very small scale, it also changed our view of the availability of knowledge. Although he is universally recognized for his brilliance and the importance of his work (he received the Nobel Prize in 1932, for example), Heisenberg remained in Germany during World War II and headed the German effort to build a nuclear bomb, permanently alienating himself from most of the scientific community. (credit: Author Unknown, via Wikimedia Commons)

It was Werner Heisenberg who first stated this limit to knowledge in 1929 as a result of his work on quantum mechanics and the wave characteristics of all particles. (See [link]). Specifically, consider simultaneously measuring the position and momentum of an electron (it could be any particle). There is an **uncertainty in position** Δx that is approximately equal to the wavelength of the particle. That is,

Equation:

$$\Delta x \approx \lambda.$$

As discussed above, a wave is not located at one point in space. If the electron's position is measured repeatedly, a spread in locations will be observed, implying an uncertainty in position Δx. To detect the position of the particle, we must interact with it, such as having it collide with a detector. In the collision, the particle will lose momentum. This change in momentum could be anywhere from close to zero to the total momentum of the particle, $p = h/\lambda$. It is not possible to tell how much momentum will be transferred to a detector, and so there is an **uncertainty in momentum** Δp, too. In fact, the uncertainty in momentum may be as large as the momentum itself, which in equation form means that

Equation:

$$\Delta p \approx \frac{h}{\lambda}.$$

The uncertainty in position can be reduced by using a shorter-wavelength electron, since $\Delta x \approx \lambda$. But shortening the wavelength increases the uncertainty in momentum, since $\Delta p \approx h/\lambda$. Conversely, the uncertainty in momentum can be reduced by using a longer-wavelength electron, but this increases the uncertainty in position. Mathematically, you can express this trade-off by multiplying the uncertainties. The wavelength cancels, leaving

Equation:

$$\Delta x \Delta p \approx h.$$

So if one uncertainty is reduced, the other must increase so that their product is $\approx h$.

With the use of advanced mathematics, Heisenberg showed that the best that can be done in a *simultaneous measurement of position and momentum* is

Equation:

$$\Delta x \Delta p \geq \frac{h}{4\pi}.$$

This is known as the **Heisenberg uncertainty principle**. It is impossible to measure position x and momentum p simultaneously with uncertainties Δx and Δp that multiply to be less than $h/4\pi$. Neither uncertainty can be zero. Neither uncertainty can become small without the other becoming large. A small wavelength allows accurate position measurement, but it increases the momentum of the probe to the point that it further disturbs the momentum of a system being measured. For example, if an electron is scattered from an atom and has a wavelength small enough to detect the position of electrons in the atom, its momentum can knock the electrons from their orbits in a manner that loses information about their original motion. It is therefore impossible to follow an electron in its orbit around an atom. If you measure the electron's position, you will find it in a definite location, but the atom will be disrupted. Repeated measurements on identical atoms will produce interesting probability distributions for electrons around the atom, but they will not produce motion information. The probability distributions are referred to as electron clouds or orbitals. The shapes of these orbitals are often shown in general chemistry texts and are discussed in The Wave Nature of Matter Causes Quantization.

Example:
Heisenberg Uncertainty Principle in Position and Momentum for an Atom

(a) If the position of an electron in an atom is measured to an accuracy of 0.0100 nm, what is the electron's uncertainty in velocity? (b) If the electron has this velocity, what is its kinetic energy in eV?

Strategy

The uncertainty in position is the accuracy of the measurement, or $\Delta x = 0.0100$ nm. Thus the smallest uncertainty in momentum Δp can be calculated using $\Delta x \Delta p \geq h/4\pi$. Once the uncertainty in momentum Δp is found, the uncertainty in velocity can be found from $\Delta p = m\Delta v$.

Solution for (a)

Using the equals sign in the uncertainty principle to express the minimum uncertainty, we have

Equation:

$$\Delta x \Delta p = \frac{h}{4\pi}.$$

Solving for Δp and substituting known values gives

Equation:

$$\Delta p = \frac{h}{4\pi\Delta x} = \frac{6.63 \times 10^{-34}\ \text{J} \cdot \text{s}}{4\pi(1.00 \times 10^{-11}\ \text{m})} = 5.28 \times 10^{-24}\ \text{kg} \cdot \text{m/s}.$$

Thus,

Equation:

$$\Delta p = 5.28 \times 10^{-24}\ \text{kg} \cdot \text{m/s} = m\Delta v.$$

Solving for Δv and substituting the mass of an electron gives

Equation:

$$\Delta v = \frac{\Delta p}{m} = \frac{5.28 \times 10^{-24}\ \text{kg} \cdot \text{m/s}}{9.11 \times 10^{-31}\ \text{kg}} = 5.79 \times 10^{6}\ \text{m/s}.$$

Solution for (b)

Although large, this velocity is not highly relativistic, and so the electron’s kinetic energy is

Equation:

$$\begin{aligned} \text{KE}_e &= \tfrac{1}{2}mv^2 \\ &= \tfrac{1}{2}(9.11 \times 10^{-31}\ \text{kg})(5.79 \times 10^{6}\ \text{m/s})^2 \\ &= (1.53 \times 10^{-17}\ \text{J})\left(\frac{1\ \text{eV}}{1.60 \times 10^{-19}\ \text{J}}\right) = 95.5\ \text{eV}. \end{aligned}$$

Discussion

Since atoms are roughly 0.1 nm in size, knowing the position of an electron to 0.0100 nm localizes it reasonably well inside the atom. This would be like being able to see details one-tenth the size of the atom. But the consequent uncertainty in velocity is large. You certainly could not follow it very well if its velocity is so uncertain. To get a further idea of how large the uncertainty in velocity is, we assumed the velocity of the electron was equal to its uncertainty and found this gave a kinetic energy of 95.5 eV. This is significantly greater than the typical energy difference between levels in atoms (see [link]), so that it is impossible to get a meaningful energy for the electron if we know its position even moderately well.

Why don’t we notice Heisenberg’s uncertainty principle in everyday life? The answer is that Planck’s constant is very small. Thus the lower limit in the uncertainty of measuring the position and momentum of large objects is negligible. We can detect sunlight reflected from Jupiter and follow the planet in its orbit around the Sun. The reflected sunlight alters the momentum of Jupiter and creates an uncertainty in its momentum, but this is totally negligible compared with Jupiter’s huge momentum. The correspondence principle tells us that the predictions of quantum mechanics become indistinguishable from classical physics for large objects, which is the case here.

Heisenberg Uncertainty for Energy and Time

There is another form of **Heisenberg’s uncertainty principle** for *simultaneous measurements of energy and time*. In equation form,
Equation:

$$\Delta E \Delta t \geq \frac{h}{4\pi},$$

where ΔE is the **uncertainty in energy** and Δt is the **uncertainty in time**. This means that within a time interval Δt, it is not possible to measure energy precisely—there will be an uncertainty ΔE in the measurement. In order to measure energy more precisely (to make ΔE smaller), we must increase Δt. This time interval may be the amount of time we take to make the measurement, or it could be the amount of time a particular state exists, as in the next [link].

Example:

Heisenberg Uncertainty Principle for Energy and Time for an Atom

An atom in an excited state temporarily stores energy. If the lifetime of this excited state is measured to be 1.0×10^{-10} s, what is the minimum uncertainty in the energy of the state in eV?

Strategy

The minimum uncertainty in energy ΔE is found by using the equals sign in $\Delta E\Delta t \geq h/4\pi$ and corresponds to a reasonable choice for the uncertainty in time. The largest the uncertainty in time can be is the full lifetime of the excited state, or $\Delta t = 1.0\times10^{-10}$ s.

Solution

Solving the uncertainty principle for ΔE and substituting known values gives

Equation:

$$\Delta E = \frac{h}{4\pi\Delta \mathrm{t}} = \frac{6.63\times10^{-34}\text{ J}\cdot\text{s}}{4\pi(1.0\times10^{-10}\text{ s})} = 5.3\times10^{-25}\text{ J.}$$

Now converting to eV yields

Equation:

$$\Delta E = (5.3\times10^{-25}\text{ J})\ \frac{1\text{ eV}}{1.6\times10^{-19}\text{ J}} = 3.3\times10^{-6}\text{ eV.}$$

Discussion

The lifetime of 10^{-10} s is typical of excited states in atoms—on human time scales, they quickly emit their stored energy. An uncertainty in energy of only a few millionths of an eV results. This uncertainty is small compared with typical excitation energies in atoms, which are on the order of 1 eV. So here the uncertainty principle limits the accuracy with which we can measure the lifetime and energy of such states, but not very significantly.

The uncertainty principle for energy and time can be of great significance if the lifetime of a system is very short. Then Δt is very small, and ΔE is consequently very large. Some nuclei and exotic particles have extremely short lifetimes (as small as 10^{-25} s), causing uncertainties in energy as great as many GeV (10^9 eV). Stored energy appears as increased rest mass, and so this means that there is significant uncertainty in the rest mass of short-lived particles. When measured repeatedly, a spread of masses or decay energies are obtained. The spread is ΔE. You might ask whether this uncertainty in energy could be avoided by not measuring the lifetime. The answer is no. Nature knows the lifetime, and so its brevity affects the energy of the particle. This is so well established experimentally that the uncertainty in decay energy is used to calculate the lifetime of short-lived states. Some nuclei and particles are so short-lived that it is difficult to measure their lifetime. But if their decay energy can be measured, its spread is ΔE, and this is used in the uncertainty principle ($\Delta E \Delta t \geq h/4\pi$) to calculate the lifetime Δt.

There is another consequence of the uncertainty principle for energy and time. If energy is uncertain by ΔE, then conservation of energy can be violated by ΔE for a time Δt. Neither the physicist nor nature can tell that conservation of energy has been violated, if the violation is temporary and smaller than the uncertainty in energy. While this sounds innocuous enough, we shall see in later chapters that it allows the temporary creation of matter from nothing and has implications for how nature transmits forces over very small distances.

Finally, note that in the discussion of particles and waves, we have stated that individual measurements produce precise or particle-like results. A definite position is determined each time we observe an electron, for example. But repeated measurements produce a spread in values consistent with wave characteristics. The great theoretical physicist Richard Feynman (1918–1988) commented, “What there are, are particles.” When you observe enough of them, they distribute themselves as you would expect for a wave phenomenon. However, what there are as they travel we cannot tell because, when we do try to measure, we affect the traveling.

Section Summary

- Matter is found to have the same interference characteristics as any other wave.
- There is now a probability distribution for the location of a particle rather than a definite position.
- Another consequence of the wave character of all particles is the Heisenberg uncertainty principle, which limits the precision with which certain physical quantities can be known simultaneously. For position and momentum, the uncertainty principle is $\Delta x \Delta p \geq \frac{h}{4\pi}$, where Δx is the uncertainty in position and Δp is the uncertainty in momentum.
- For energy and time, the uncertainty principle is $\Delta E \Delta t \geq \frac{h}{4\pi}$ where ΔE is the uncertainty in energy and Δt is the uncertainty in time.
- These small limits are fundamentally important on the quantum-mechanical scale.

Conceptual Questions

Exercise:

Problem:

What is the Heisenberg uncertainty principle? Does it place limits on what can be known?

Problems & Exercises

Exercise:

Problem:

(a) If the position of an electron in a membrane is measured to an accuracy of $1.00\ \mu\text{m}$, what is the electron's minimum uncertainty in velocity? (b) If the electron has this velocity, what is its kinetic energy in eV? (c) What are the implications of this energy, comparing it to typical molecular binding energies?

Solution:

(a) 57.9 m/s

(b) $9.55 \times 10^{-9}\ \text{eV}$

(c) From [link], we see that typical molecular binding energies range from about 1eV to 10 eV, therefore the result in part (b) is approximately 9 orders of magnitude smaller than typical molecular binding energies.

Exercise:

Problem:

(a) If the position of a chlorine ion in a membrane is measured to an accuracy of $1.00\ \mu\text{m}$, what is its minimum uncertainty in velocity, given its mass is $5.86 \times 10^{-26}\ \text{kg}$? (b) If the ion has this velocity, what is its kinetic energy in eV, and how does this compare with typical molecular binding energies?

Exercise:

Problem:

Suppose the velocity of an electron in an atom is known to an accuracy of $2.0 \times 10^{3}\ \text{m/s}$ (reasonably accurate compared with orbital velocities). What is the electron's minimum uncertainty in position, and how does this compare with the approximate 0.1-nm size of the atom?

Solution:

29 nm,

290 times greater

Exercise:

Problem:

The velocity of a proton in an accelerator is known to an accuracy of 0.250% of the speed of light. (This could be small compared with its velocity.) What is the smallest possible uncertainty in its position?

Exercise:

Problem:

A relatively long-lived excited state of an atom has a lifetime of 3.00 ms. What is the minimum uncertainty in its energy?

Solution:

1.10×10^{-13} eV

Exercise:

Problem:

(a) The lifetime of a highly unstable nucleus is 10^{-20} s. What is the smallest uncertainty in its decay energy? (b) Compare this with the rest energy of an electron.

Exercise:

Problem:

The decay energy of a short-lived particle has an uncertainty of 1.0 MeV due to its short lifetime. What is the smallest lifetime it can have?

Solution:

3.3×10^{-22} s

Exercise:

Problem:

The decay energy of a short-lived nuclear excited state has an uncertainty of 2.0 eV due to its short lifetime. What is the smallest lifetime it can have?

Exercise:

Problem:

What is the approximate uncertainty in the mass of a muon, as determined from its decay lifetime?

Solution:

2.66×10^{-46} kg

Exercise:

Problem:

Derive the approximate form of Heisenberg’s uncertainty principle for energy and time, $\Delta E \Delta t \approx h$, using the following arguments: Since the position of a particle is uncertain by $\Delta x \approx \lambda$, where λ is the wavelength of the photon used to examine it, there is an uncertainty in the time the photon takes to traverse Δx. Furthermore, the photon has an energy related to its wavelength, and it can transfer some or all of this energy to the object being examined. Thus the uncertainty in the energy of the object is also related to λ. Find Δt and ΔE; then multiply them to give the approximate uncertainty principle.

Glossary

Heisenberg’s uncertainty principle
: a fundamental limit to the precision with which pairs of quantities (momentum and position, and energy and time) can be measured

uncertainty in energy

lack of precision or lack of knowledge of precise results in measurements of energy

uncertainty in time

lack of precision or lack of knowledge of precise results in measurements of time

uncertainty in momentum

lack of precision or lack of knowledge of precise results in measurements of momentum

uncertainty in position

lack of precision or lack of knowledge of precise results in measurements of position

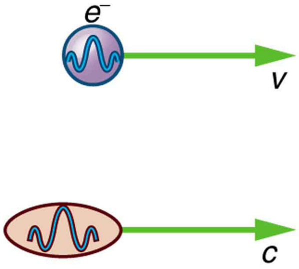
probability distribution

the overall spatial distribution of probabilities to find a particle at a given location

The Particle-Wave Duality Reviewed

- Explain the concept of particle-wave duality, and its scope.

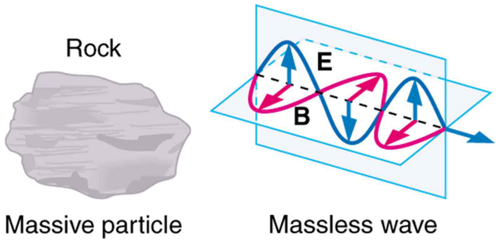
Particle-wave duality—the fact that all particles have wave properties—is one of the cornerstones of quantum mechanics. We first came across it in the treatment of photons, those particles of EM radiation that exhibit both particle and wave properties, but not at the same time. Later it was noted that particles of matter have wave properties as well. The dual properties of particles and waves are found for all particles, whether massless like photons, or having a mass like electrons. (See [link].)



On a quantum-mechanical scale (i.e., very small), particles with and without mass have wave properties. For example, both electrons and photons have wavelengths but also behave as particles.

There are many submicroscopic particles in nature. Most have mass and are expected to act as particles, or the smallest units of matter. All these masses have wave properties, with wavelengths given by the de Broglie relationship $\lambda = h/p$. So, too, do combinations of these particles, such as nuclei, atoms, and molecules. As a combination of masses becomes large, particularly if it is large enough to be called macroscopic, its wave nature becomes difficult to observe. This is consistent with our common experience with matter.

Some particles in nature are massless. We have only treated the photon so far, but all massless entities travel at the speed of light, have a wavelength, and exhibit particle and wave behaviors. They have momentum given by a rearrangement of the de Broglie relationship, $p = h/\lambda$. In large combinations of these massless particles (such large combinations are common only for photons or EM waves), there is mostly wave behavior upon detection, and the particle nature becomes difficult to observe. This is also consistent with experience. (See [link].)



On a classical scale (macroscopic), particles with mass behave as particles and not as waves. Particles without mass act as waves and not as particles.

The particle-wave duality is a universal attribute. It is another connection between matter and energy. Not only has modern physics been able to describe nature for high speeds and small sizes, it has also discovered new connections and symmetries. There is greater unity and symmetry in nature than was known in the classical era—but they were dreamt of. A beautiful poem written by the English poet William Blake some two centuries ago contains the following four lines:

To see the World in a Grain of Sand

And a Heaven in a Wild Flower

Hold Infinity in the palm of your hand

And Eternity in an hour

Integrated Concepts

The problem set for this section involves concepts from this chapter and several others. Physics is most interesting when applied to general situations involving more than a narrow set of physical principles. For example, photons have momentum, hence the relevance of Linear Momentum and Collisions. The following topics are involved in some or all of the problems in this section:

- Dynamics: Newton’s Laws of Motion
- Work, Energy, and Energy Resources
- Linear Momentum and Collisions
- Heat and Heat Transfer Methods
- Electric Potential and Electric Field
- Electric Current, Resistance, and Ohm’s Law
- Wave Optics
- Special Relativity

Note:
Problem-Solving Strategy

1. Identify which physical principles are involved.

2. Solve the problem using strategies outlined in the text.

[link] illustrates how these strategies are applied to an integrated-concept problem.

Example:
Recoil of a Dust Particle after Absorbing a Photon
The following topics are involved in this integrated concepts worked example:

Photons (quantum mechanics)
Linear Momentum

Topics

A 550-nm photon (visible light) is absorbed by a 1.00-μg particle of dust in outer space. (a) Find the momentum of such a photon. (b) What is the recoil velocity of the particle of dust, assuming it is initially at rest?

Strategy Step 1

To solve an *integrated-concept problem*, such as those following this example, we must first identify the physical principles involved and identify the chapters in which they are found. Part (a) of this example asks for the *momentum of a photon*, a topic of the present chapter. Part (b) considers *recoil following a collision*, a topic of Linear Momentum and Collisions.

Strategy Step 2

The following solutions to each part of the example illustrate how specific problem-solving strategies are applied. These involve identifying knowns and unknowns, checking to see if the answer is reasonable, and so on.

Solution for (a)

The momentum of a photon is related to its wavelength by the equation:

Equation:

$$p=\frac{h}{\lambda}.$$

Entering the known value for Planck's constant h and given the wavelength λ, we obtain

Equation:

$$\begin{array}{lll} p & = & \frac{6.63\times 10^{-34}\text{ J}\cdot\text{s}}{550\times 10^{-9}\text{ m}} \\ & = & 1.21\times 10^{-27}\text{ kg}\cdot\text{m/s}. \end{array}$$

Discussion for (a)

This momentum is small, as expected from discussions in the text and the fact that photons of visible light carry small amounts of energy and momentum compared with those carried by macroscopic objects.

Solution for (b)

Conservation of momentum in the absorption of this photon by a grain of dust can be analyzed using the equation:

Equation:

$$p_1+p_2=p\prime_1+p\prime_2(F_{\text{net}}=0).$$

The net external force is zero, since the dust is in outer space. Let 1 represent the photon and 2 the dust particle. Before the collision, the dust is at rest (relative to some observer); after the collision, there is no photon (it is absorbed). So conservation of momentum can be written

Equation:

$$p_1=p\prime_2=\text{mv},$$

where p_1 is the photon momentum before the collision and $p\prime_2$ is the dust momentum after the collision. The mass and recoil velocity of the dust are

m and v, respectively. Solving this for v, the requested quantity, yields

Equation:

$$v = \frac{p}{m},$$

where p is the photon momentum found in part (a). Entering known values (noting that a microgram is 10^{-9} kg) gives

Equation:

$$\begin{aligned} v &= \frac{1.21\times10^{-27}\text{ kg}\cdot\text{m/s}}{1.00\times10^{-9}\text{ kg}} \\ &= 1.21\times10^{-18}\text{ m/s}. \end{aligned}$$

Discussion

The recoil velocity of the particle of dust is extremely small. As we have noted, however, there are immense numbers of photons in sunlight and other macroscopic sources. In time, collisions and absorption of many photons could cause a significant recoil of the dust, as observed in comet tails.

Section Summary

- The particle-wave duality refers to the fact that all particles—those with mass and those without mass—have wave characteristics.
- This is a further connection between mass and energy.

Conceptual Questions

Exercise:

Problem:

In what ways are matter and energy related that were not known before the development of relativity and quantum mechanics?

Problems & Exercises

Exercise:

Problem: Integrated Concepts

The 54.0-eV electron in [link] has a 0.167-nm wavelength. If such electrons are passed through a double slit and have their first maximum at an angle of 25.0^{o}, what is the slit separation d?

Solution:

0.395 nm

Exercise:

Problem: Integrated Concepts

An electron microscope produces electrons with a 2.00-pm wavelength. If these are passed through a 1.00-nm single slit, at what angle will the first diffraction minimum be found?

Exercise:

Problem: Integrated Concepts

A certain heat lamp emits 200 W of mostly IR radiation averaging 1500 nm in wavelength. (a) What is the average photon energy in joules? (b) How many of these photons are required to increase the temperature of a person's shoulder by 2.0^{o}C, assuming the affected mass is 4.0 kg with a specific heat of $0.83\ \text{kcal/kg}\cdot^{\text{o}}\text{C}$. Also assume no other significant heat transfer. (c) How long does this take?

Solution:

(a) $1.3\times 10^{-19}\ \text{J}$

(b) 2.1×10^{23}

(c) $1.4 \times 10^2\ \text{s}$

Exercise:

Problem: Integrated Concepts

On its high power setting, a microwave oven produces 900 W of 2560 MHz microwaves. (a) How many photons per second is this? (b) How many photons are required to increase the temperature of a 0.500-kg mass of pasta by 45.0^{o}C, assuming a specific heat of $0.900\ \text{kcal/kg} \cdot {}^{\text{o}}\text{C}$? Neglect all other heat transfer. (c) How long must the microwave operator wait for their pasta to be ready?

Exercise:

Problem: Integrated Concepts

(a) Calculate the amount of microwave energy in joules needed to raise the temperature of 1.00 kg of soup from 20.0^{o}C to 100^{o}C. (b) What is the total momentum of all the microwave photons it takes to do this? (c) Calculate the velocity of a 1.00-kg mass with the same momentum. (d) What is the kinetic energy of this mass?

Solution:

(a) $3.35 \times 10^5\ \text{J}$

(b) $1.12 \times 10^{-3}\ \text{kg} \cdot \text{m/s}$

(c) $1.12 \times 10^{-3}\ \text{m/s}$

(d) $6.23 \times 10^{-7}\ \text{J}$

Exercise:

Problem: Integrated Concepts

(a) What is γ for an electron emerging from the Stanford Linear Accelerator with a total energy of 50.0 GeV? (b) Find its momentum. (c) What is the electron's wavelength?

Exercise:

Problem: Integrated Concepts

(a) What is γ for a proton having an energy of 1.00 TeV, produced by the Fermilab accelerator? (b) Find its momentum. (c) What is the proton's wavelength?

Solution:

(a) 1.06×10^{3}

(b) $5.33 \times 10^{-16}\ \text{kg} \cdot \text{m/s}$

(c) $1.24 \times 10^{-18}\ \text{m}$

Exercise:

Problem: Integrated Concepts

An electron microscope passes 1.00-pm-wavelength electrons through a circular aperture $2.00\ \mu\text{m}$ in diameter. What is the angle between two just-resolvable point sources for this microscope?

Exercise:

Problem: Integrated Concepts

(a) Calculate the velocity of electrons that form the same pattern as 450-nm light when passed through a double slit. (b) Calculate the kinetic energy of each and compare them. (c) Would either be easier to generate than the other? Explain.

Solution:

(a) $1.62 \times 10^{3}\ \text{m/s}$

(b) 4.42×10^{-19} J for photon, 1.19×10^{-24} J for electron, photon energy is 3.71×10^{5} times greater

(c) The light is easier to make because 450-nm light is blue light and therefore easy to make. Creating electrons with $7.43\ \mu\text{eV}$ of energy would not be difficult, but would require a vacuum.

Exercise:

Problem: Integrated Concepts

(a) What is the separation between double slits that produces a second-order minimum at 45.0° for 650-nm light? (b) What slit separation is needed to produce the same pattern for 1.00-keV protons.

Solution:

(a) $2.30 \times 10^{-6}\ \text{m}$

(b) $3.20 \times 10^{-12}\ \text{m}$

Exercise:

Problem: Integrated Concepts

A laser with a power output of 2.00 mW at a wavelength of 400 nm is projected onto calcium metal. (a) How many electrons per second are ejected? (b) What power is carried away by the electrons, given that the binding energy is 2.71 eV? (c) Calculate the current of ejected electrons. (d) If the photoelectric material is electrically insulated and acts like a 2.00-pF capacitor, how long will current flow before the capacitor voltage stops it?

Exercise:

Problem: Integrated Concepts

One problem with x rays is that they are not sensed. Calculate the temperature increase of a researcher exposed in a few seconds to a nearly fatal accidental dose of x rays under the following conditions. The energy of the x-ray photons is 200 keV, and 4.00×10^{13} of them are absorbed per kilogram of tissue, the specific heat of which is $0.830\ \text{kcal/kg} \cdot \text{ºC}$. (Note that medical diagnostic x-ray machines *cannot* produce an intensity this great.)

Solution:

$3.69 \times 10^{-4}\ \text{ºC}$

Exercise:

Problem: Integrated Concepts

A 1.00-fm photon has a wavelength short enough to detect some information about nuclei. (a) What is the photon momentum? (b) What is its energy in joules and MeV? (c) What is the (relativistic) velocity of an electron with the same momentum? (d) Calculate the electron’s kinetic energy.

Exercise:

Problem: Integrated Concepts

The momentum of light is exactly reversed when reflected straight back from a mirror, assuming negligible recoil of the mirror. Thus the change in momentum is twice the photon momentum. Suppose light of intensity $1.00\ \text{kW/m}^2$ reflects from a mirror of area $2.00\ \text{m}^2$. (a) Calculate the energy reflected in 1.00 s. (b) What is the momentum imparted to the mirror? (c) Using the most general form of Newton’s second law, what is the force on the mirror? (d) Does the assumption of no mirror recoil seem reasonable?

Solution:

(a) 2.00 kJ

(b) $1.33 \times 10^{-5}\ \text{kg} \cdot \text{m/s}$

(c) $1.33 \times 10^{-5}\ \text{N}$

(d) yes

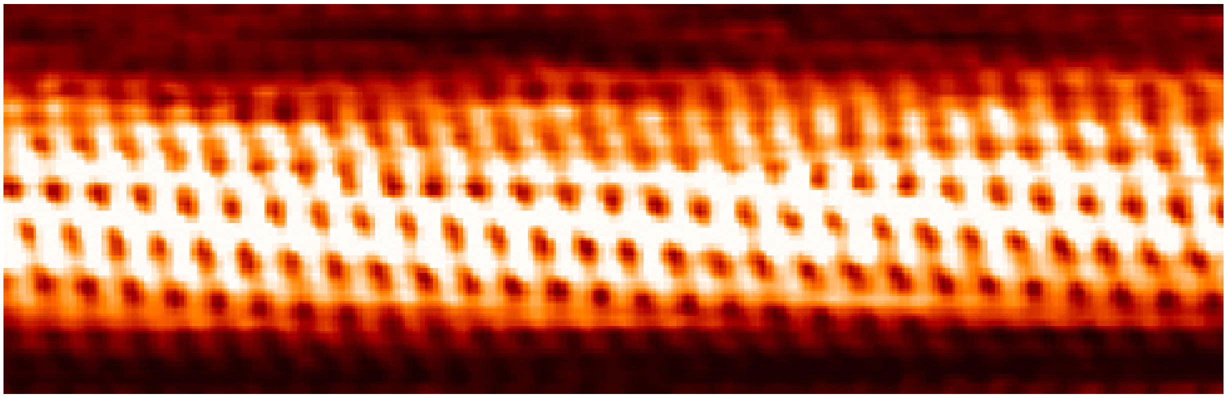
Exercise:

Problem: Integrated Concepts

Sunlight above the Earth's atmosphere has an intensity of $1.30\ \text{kW/m}^2$. If this is reflected straight back from a mirror that has only a small recoil, the light's momentum is exactly reversed, giving the mirror twice the incident momentum. (a) Calculate the force per square meter of mirror. (b) Very low mass mirrors can be constructed in the near weightlessness of space, and attached to a spaceship to sail it. Once done, the average mass per square meter of the spaceship is 0.100 kg. Find the acceleration of the spaceship if all other forces are balanced. (c) How fast is it moving 24 hours later?

Introduction to Atomic Physics
class="introduction"

Individual carbon atoms are visible in this image of a carbon nanotube made by a scanning tunneling electron microscope. (credit: Taner Yildirim, National Institute of Standards and Technology, via Wikimedia Commons)



From childhood on, we learn that atoms are a substructure of all things around us, from the air we breathe to the autumn leaves that blanket a forest trail. Invisible to the eye, the existence and properties of atoms are used to explain many phenomena—a theme found throughout this text. In this chapter, we discuss the discovery of atoms and their own substructures; we then apply quantum mechanics to the description of atoms, and their properties and interactions. Along the way, we will find, much like the scientists who made the original discoveries, that new concepts emerge with applications far beyond the boundaries of atomic physics.

Discovery of the Atom

- Describe the basic structure of the atom, the substructure of all matter.

How do we know that atoms are really there if we cannot see them with our eyes? A brief account of the progression from the proposal of atoms by the Greeks to the first direct evidence of their existence follows.

People have long speculated about the structure of matter and the existence of atoms. The earliest significant ideas to survive are due to the ancient Greeks in the fifth century BCE, especially those of the philosophers Leucippus and Democritus. (There is some evidence that philosophers in both India and China made similar speculations, at about the same time.) They considered the question of whether a substance can be divided without limit into ever smaller pieces. There are only a few possible answers to this question. One is that infinitesimally small subdivision is possible. Another is what Democritus in particular believed—that there is a smallest unit that cannot be further subdivided. Democritus called this the **atom**. We now know that atoms themselves can be subdivided, but their identity is destroyed in the process, so the Greeks were correct in a respect. The Greeks also felt that atoms were in constant motion, another correct notion.

The Greeks and others speculated about the properties of atoms, proposing that only a few types existed and that all matter was formed as various combinations of these types. The famous proposal that the basic elements were earth, air, fire, and water was brilliant, but incorrect. The Greeks had identified the most common examples of the four states of matter (solid, gas, plasma, and liquid), rather than the basic elements. More than 2000 years passed before observations could be made with equipment capable of revealing the true nature of atoms.

Over the centuries, discoveries were made regarding the properties of substances and their chemical reactions. Certain systematic features were recognized, but similarities between common and rare elements resulted in efforts to transmute them (lead into gold, in particular) for financial gain. Secrecy was endemic. Alchemists discovered and rediscovered many facts but did not make them broadly available. As the Middle Ages ended, alchemy gradually faded, and the science of chemistry arose. It was no longer possible, nor considered desirable, to keep discoveries secret. Collective knowledge grew, and by the beginning of the 19th century, an important fact was well established—the masses of reactants in specific chemical reactions always have a particular mass ratio. This is very strong indirect evidence that there are basic units (atoms and molecules) that have these same mass ratios. The English chemist John Dalton (1766–1844) did much of this work, with significant contributions by the Italian physicist Amedeo Avogadro (1776–1856). It was Avogadro who developed the idea of a fixed number of atoms and molecules in a mole, and this special number is called Avogadro’s number in his honor. The Austrian physicist Johann Josef Loschmidt was the first to measure the value of the constant in 1865 using the kinetic theory of gases.

Note:

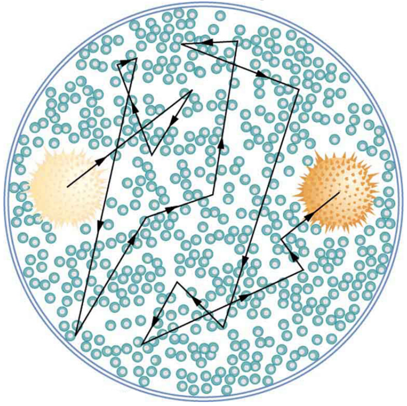
Patterns and Systematics

The recognition and appreciation of patterns has enabled us to make many discoveries. The periodic table of elements was proposed as an organized summary of the known elements long before all elements had been discovered, and it led to many other discoveries. We shall see in later chapters that patterns in the properties of subatomic particles led to the proposal of quarks as their underlying structure, an idea that is still bearing fruit.

Knowledge of the properties of elements and compounds grew, culminating in the mid-19th-century development of the periodic table of the elements by Dmitri Mendeleev (1834–1907), the great Russian chemist. Mendeleev proposed an ingenious array that highlighted the periodic nature of the properties of elements. Believing in the systematics of the periodic table, he also predicted the existence of then-unknown elements to complete it. Once these elements were discovered and determined to have properties predicted by Mendeleev, his periodic table became universally accepted.

Also during the 19th century, the kinetic theory of gases was developed. Kinetic theory is based on the existence of atoms and molecules in random thermal motion and provides a microscopic explanation of the gas laws, heat transfer, and thermodynamics (see Introduction to Temperature, Kinetic Theory, and the Gas Laws and Introduction to Laws of Thermodynamics). Kinetic theory works so well that it is another strong indication of the existence of atoms. But it is still indirect evidence—individual atoms and molecules had not been observed. There were heated debates about the validity of kinetic theory until direct evidence of atoms was obtained.

The first truly direct evidence of atoms is credited to Robert Brown, a Scottish botanist. In 1827, he noticed that tiny pollen grains suspended in still water moved about in complex paths. This can be observed with a microscope for any small particles in a fluid. The motion is caused by the random thermal motions of fluid molecules colliding with particles in the fluid, and it is now called **Brownian motion**. (See [link].) Statistical fluctuations in the numbers of molecules striking the sides of a visible particle cause it to move first this way, then that. Although the molecules cannot be directly observed, their effects on the particle can be. By examining Brownian motion, the size of molecules can be calculated. The smaller and more numerous they are, the smaller the fluctuations in the numbers striking different sides.



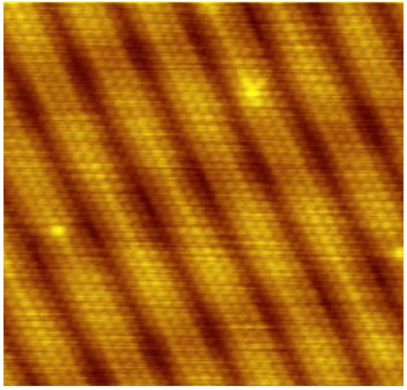
The position of a pollen grain in water, measured every few seconds under a microscope, exhibits Brownian motion. Brownian motion is due to fluctuations in the number of atoms and molecules colliding with a small mass, causing it to move about in complex paths. This is nearly direct evidence for the existence of atoms, providing a satisfactory alternative explanation cannot be found.

It was Albert Einstein who, starting in his epochal year of 1905, published several papers that explained precisely how Brownian motion could be used to measure the size of atoms and molecules. (In 1905 Einstein created special relativity, proposed photons as quanta of EM radiation, and produced a theory of Brownian motion that allowed the size of atoms to be determined. All of this was done in his spare time, since he worked days as a patent examiner. Any one of these very basic works could have been the crowning achievement of an entire career—yet Einstein did even more in later years.) Their sizes were only approximately known to be 10^{-10} m, based on a comparison of latent heat of vaporization and surface tension made in about 1805 by Thomas Young of double-slit fame and the famous astronomer and mathematician Simon Laplace.

Using Einstein’s ideas, the French physicist Jean-Baptiste Perrin (1870–1942) carefully observed Brownian motion; not only did he confirm Einstein’s theory, he also produced accurate sizes for atoms and molecules.

Since molecular weights and densities of materials were well established, knowing atomic and molecular sizes allowed a precise value for Avogadro’s number to be obtained. (If we know how big an atom is, we know how many fit into a certain volume.) Perrin also used these ideas to explain atomic and molecular agitation effects in sedimentation, and he received the 1926 Nobel Prize for his achievements. Most scientists were already convinced of the existence of atoms, but the accurate observation and analysis of Brownian motion was conclusive—it was the first truly direct evidence.

A huge array of direct and indirect evidence for the existence of atoms now exists. For example, it has become possible to accelerate ions (much as electrons are accelerated in cathode-ray tubes) and to detect them individually as well as measure their masses (see More Applications of Magnetism for a discussion of mass spectrometers). Other devices that observe individual atoms, such as the scanning tunneling electron microscope, will be discussed elsewhere. (See [link].) All of our understanding of the properties of matter is based on and consistent with the atom. The atom’s substructures, such as electron shells and the nucleus, are both interesting and important. The nucleus in turn has a substructure, as do the particles of which it is composed. These topics, and the question of whether there is a smallest basic structure to matter, will be explored in later parts of the text.



Individual atoms can be detected with devices such as the scanning tunneling electron microscope that produced this image of individual gold atoms on a graphite substrate. (credit: Erwin Rossen, Eindhoven University of Technology, via Wikimedia Commons)

Section Summary

- Atoms are the smallest unit of elements; atoms combine to form molecules, the smallest unit of compounds.
- The first direct observation of atoms was in Brownian motion.
- Analysis of Brownian motion gave accurate sizes for atoms (10^{-10} m on average) and a precise value for Avogadro's number.

Conceptual Questions

Exercise:

Problem:

Name three different types of evidence for the existence of atoms.

Exercise:

Problem:

Explain why patterns observed in the periodic table of the elements are evidence for the existence of atoms, and why Brownian motion is a more direct type of evidence for their existence.

Exercise:

Problem: If atoms exist, why can’t we see them with visible light?

Problems & Exercises

Exercise:

Problem:

Using the given charge-to-mass ratios for electrons and protons, and knowing the magnitudes of their charges are equal, what is the ratio of the proton’s mass to the electron’s? (Note that since the charge-to-mass ratios are given to only three-digit accuracy, your answer may differ from the accepted ratio in the fourth digit.)

Solution:

1.84×10^3

Exercise:

Problem:

(a) Calculate the mass of a proton using the charge-to-mass ratio given for it in this chapter and its known charge. (b) How does your result compare with the proton mass given in this chapter?

Exercise:

Problem:

If someone wanted to build a scale model of the atom with a nucleus 1.00 m in diameter, how far away would the nearest electron need to be?

Solution:

50 km

Glossary

atom
: basic unit of matter, which consists of a central, positively charged nucleus surrounded by negatively charged electrons

Brownian motion
: the continuous random movement of particles of matter suspended in a liquid or gas

Discovery of the Parts of the Atom: Electrons and Nuclei

- Describe how electrons were discovered.
- Explain the Millikan oil drop experiment.
- Describe Rutherford’s gold foil experiment.
- Describe Rutherford’s planetary model of the atom.

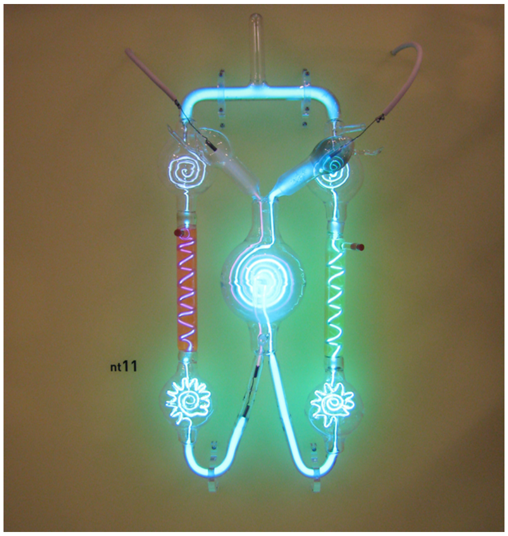
Just as atoms are a substructure of matter, electrons and nuclei are substructures of the atom. The experiments that were used to discover electrons and nuclei reveal some of the basic properties of atoms and can be readily understood using ideas such as electrostatic and magnetic force, already covered in previous chapters.

Note:
Charges and Electromagnetic Forces
In previous discussions, we have noted that positive charge is associated with nuclei and negative charge with electrons. We have also covered many aspects of the electric and magnetic forces that affect charges. We will now explore the discovery of the electron and nucleus as substructures of the atom and examine their contributions to the properties of atoms.

The Electron

Gas discharge tubes, such as that shown in [link], consist of an evacuated glass tube containing two metal electrodes and a rarefied gas. When a high voltage is applied to the electrodes, the gas glows. These tubes were the precursors to today’s neon lights. They were first studied seriously by Heinrich Geissler, a German inventor and glassblower, starting in the 1860s. The English scientist William Crookes, among others, continued to study what for some time were called Crookes tubes, wherein electrons are freed from atoms and molecules in the rarefied gas inside the tube and are accelerated from the cathode (negative) to the anode (positive) by the high potential. These “*cathode rays*” collide with the gas atoms and molecules and excite them, resulting in the emission of electromagnetic (EM) radiation that makes the electrons’ path visible as a ray that spreads and fades as it moves away from the cathode.

Gas discharge tubes today are most commonly called **cathode-ray tubes**, because the rays originate at the cathode. Crookes showed that the electrons carry momentum (they can make a small paddle wheel rotate). He also found that their normally straight path is bent by a magnet in the direction expected for a negative charge moving away from the cathode. These were the first direct indications of electrons and their charge.



A gas discharge tube glows when a high voltage is applied to it. Electrons emitted from the cathode are accelerated toward the anode; they excite atoms and molecules in the gas, which glow in response. Once called Geissler tubes and later Crookes tubes, they are now known as cathode-ray tubes (CRTs) and are found in older TVs, computer screens, and x-ray machines. When a magnetic field is applied, the beam bends in the direction expected for negative charge. (credit: Paul Downey, Flickr)

The English physicist J. J. Thomson (1856–1940) improved and expanded the scope of experiments with gas discharge tubes. (See [link] and [link].) He verified the negative charge of the cathode rays with both magnetic and electric fields. Additionally, he collected the rays in a metal cup and found an excess of negative charge. Thomson was also able to measure the ratio of the charge of the electron to its mass, q_e/m_e—an important step to finding the actual values of both q_e and m_e. [link] shows a cathode-ray tube, which produces a narrow beam of electrons that passes through charging plates connected to a high-voltage power supply. An electric field $\mathbf{E}$ is produced between the charging plates, and the cathode-ray tube is placed between the poles of a magnet so that the electric field $\mathbf{E}$ is perpendicular to the magnetic field $\mathbf{B}$ of the magnet. These fields, being perpendicular to each other, produce opposing forces on the electrons. As discussed for mass spectrometers in More Applications of Magnetism, if the net force due to the fields vanishes, then the velocity of the charged particle is $v = E/B$. In this manner, Thomson determined the velocity of the electrons and then moved the beam up and down by adjusting the electric field.

J. J. Thomson (credit: www.firstworldwar.com , via Wikimedia Commons)

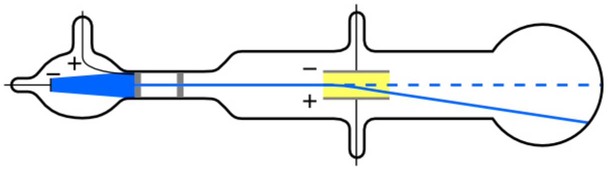
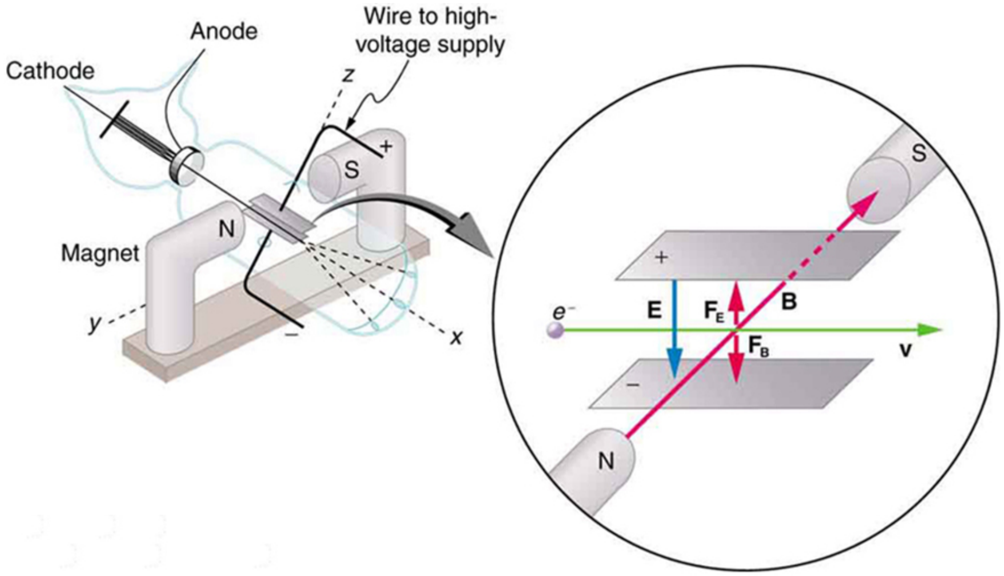


Diagram of Thomson’s CRT. (credit: Kurzon, Wikimedia Commons)



This schematic shows the electron beam in a CRT passing through crossed electric and magnetic fields and causing phosphor to glow when striking the end of the tube.

To see how the amount of deflection is used to calculate q_e/m_e, note that the deflection is proportional to the electric force on the electron:

Equation:

$$F = q_e E.$$

But the vertical deflection is also related to the electron’s mass, since the electron’s acceleration is

Equation:

$$a = \frac{F}{m_e}.$$

The value of F is not known, since q_e was not yet known. Substituting the expression for electric force into the expression for acceleration yields

Equation:

$$a=\frac{F}{m_e}=\frac{q_eE}{m_e}.$$

Gathering terms, we have
Equation:

$$\frac{q_e}{m_e}=\frac{a}{E}.$$

The deflection is analyzed to get a, and E is determined from the applied voltage and distance between the plates; thus, $\frac{q_e}{m_e}$ can be determined. With the velocity known, another measurement of $\frac{q_e}{m_e}$ can be obtained by bending the beam of electrons with the magnetic field. Since $F_{\text{mag}}=q_e\text{vB}=m_ea$, we have $q_e/m_e=a/\text{vB}$. Consistent results are obtained using magnetic deflection.

What is so important about q_e/m_e, the ratio of the electron’s charge to its mass? The value obtained is
Equation:

$$\frac{q_e}{m_e}=-1.76\times10^{11}\ \text{C/kg (electron)}.$$

This is a huge number, as Thomson realized, and it implies that the electron has a very small mass. It was known from electroplating that about $10^8\ \text{C/kg}$ is needed to plate a material, a factor of about 1000 less than the charge per kilogram of electrons. Thomson went on to do the same experiment for positively charged hydrogen ions (now known to be bare protons) and found a charge per kilogram about 1000 times smaller than that for the electron, implying that the proton is about 1000 times more massive than the electron. Today, we know more precisely that
Equation:

$$\frac{q_p}{m_p} = 9.58 \times 10^7 \text{ C/kg (proton)},$$

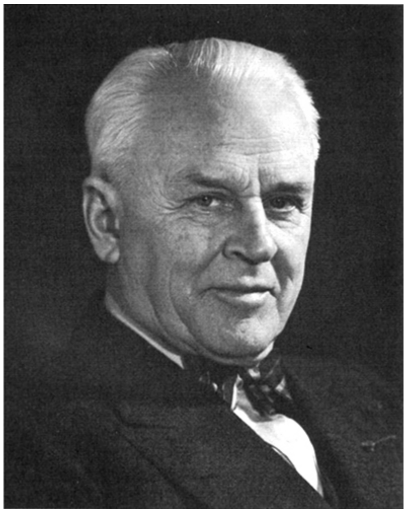
where q_p is the charge of the proton and m_p is its mass. This ratio (to four significant figures) is 1836 times less charge per kilogram than for the electron. Since the charges of electrons and protons are equal in magnitude, this implies $m_p = 1836m_e$.

Thomson performed a variety of experiments using differing gases in discharge tubes and employing other methods, such as the photoelectric effect, for freeing electrons from atoms. He always found the same properties for the electron, proving it to be an independent particle. For his work, the important pieces of which he began to publish in 1897, Thomson was awarded the 1906 Nobel Prize in Physics. In retrospect, it is difficult to appreciate how astonishing it was to find that the atom has a substructure. Thomson himself said, “It was only when I was convinced that the experiment left no escape from it that I published my belief in the existence of bodies smaller than atoms.”

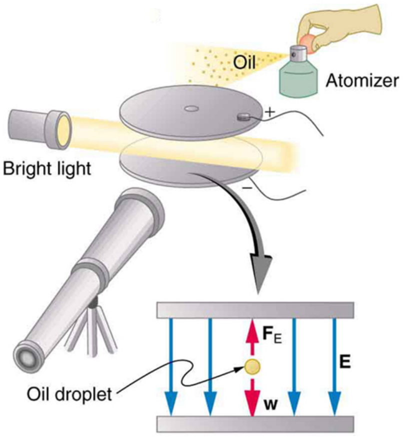
Thomson attempted to measure the charge of individual electrons, but his method could determine its charge only to the order of magnitude expected.

Since Faraday’s experiments with electroplating in the 1830s, it had been known that about 100,000 C per mole was needed to plate singly ionized ions. Dividing this by the number of ions per mole (that is, by Avogadro’s number), which was approximately known, the charge per ion was calculated to be about 1.6×10^{-19} C, close to the actual value.

An American physicist, Robert Millikan (1868–1953) (see [link]), decided to improve upon Thomson’s experiment for measuring q_e and was eventually forced to try another approach, which is now a classic experiment performed by students. The Millikan oil drop experiment is shown in [link].



Robert Millikan (credit: Unknown Author, via Wikimedia Commons)



The Millikan oil drop experiment produced the first accurate direct measurement of the charge on electrons, one of the most fundamental constants in nature. Fine drops of oil become charged when sprayed. Their movement is observed between metal plates with a potential applied to oppose the gravitational force. The balance of gravitational and electric forces allows the calculation of the charge on a drop. The charge is found to be quantized in units of -1.6×10^{-19} C, thus determining directly the charge of the excess and missing electrons on the oil drops.

In the Millikan oil drop experiment, fine drops of oil are sprayed from an atomizer. Some of these are charged by the process and can then be suspended between metal plates by a voltage between the plates. In this situation, the weight of the drop is balanced by the electric force:

Equation:

$$m_{\text{drop}}g = q_e E$$

The electric field is produced by the applied voltage, hence, $E = V/d$, and V is adjusted to just balance the drop's weight. The drops can be seen as points of reflected light using a microscope, but they are too small to directly measure their size and mass. The mass of the drop is determined by observing how fast it falls when the voltage is turned off. Since air resistance is very significant for these submicroscopic drops, the more massive drops fall faster than the less massive, and sophisticated sedimentation calculations can reveal their mass. Oil is used rather than water, because it does not readily evaporate, and so mass is nearly constant. Once the mass of the drop is known, the charge of the electron is given by rearranging the previous equation:

Equation:

$$q = \frac{m_{\text{drop}}g}{E} = \frac{m_{\text{drop}}\text{gd}}{V},$$

where d is the separation of the plates and V is the voltage that holds the drop motionless. (The same drop can be observed for several hours to see that it really is motionless.) By 1913 Millikan had measured the charge of the electron q_e to an accuracy of 1%, and he improved this by a factor of 10 within a few years to a value of -1.60×10^{-19} C. He also observed that all charges were multiples of the basic electron charge and that sudden changes could occur in which electrons were added or removed from the drops. For this very fundamental direct measurement of q_e and for his studies of the photoelectric effect, Millikan was awarded the 1923 Nobel Prize in Physics.

With the charge of the electron known and the charge-to-mass ratio known, the electron's mass can be calculated. It is

Equation:

$$m = \frac{q_e}{\left(\frac{q_e}{m_e}\right)}.$$

Substituting known values yields

Equation:

$$m_e = \frac{-1.60 \times 10^{-19}\text{ C}}{-1.76 \times 10^{11}\text{ C/kg}}$$

or

Equation:

$$m_e = 9.11 \times 10^{-31}\text{ kg} \;\; \text{(electron's mass)},$$

where the round-off errors have been corrected. The mass of the electron has been verified in many subsequent experiments and is now known to an accuracy of better than one part in one million. It is an incredibly small mass and remains the smallest known mass of any particle that has mass. (Some particles, such as photons, are massless and cannot be brought to rest, but travel at the speed of light.) A similar calculation gives the masses of other particles, including the proton. To three digits, the mass of the proton is now known to be

Equation:

$$m_p = 1.67 \times 10^{-27}\text{ kg} \;\; \text{(proton's mass)},$$

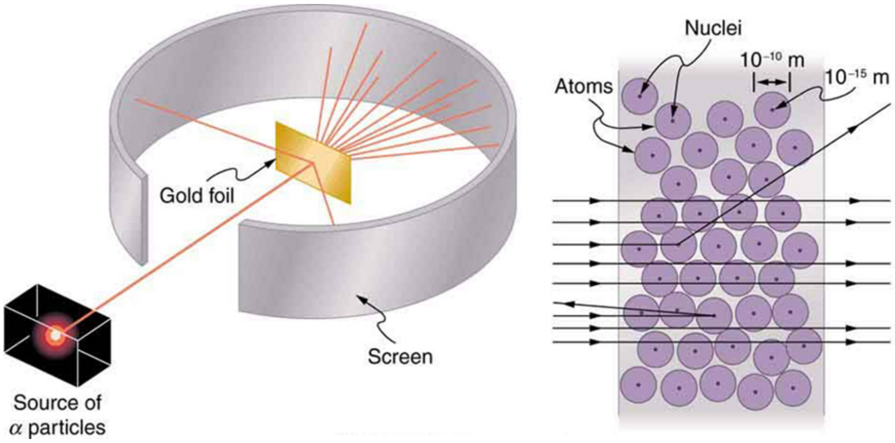
which is nearly identical to the mass of a hydrogen atom. What Thomson and Millikan had done was to prove the existence of one substructure of atoms, the electron, and further to show that it had only a tiny fraction of the mass of an atom. The nucleus of an atom contains most of its mass, and the nature of the nucleus was completely unanticipated.

Another important characteristic of quantum mechanics was also beginning to emerge. All electrons are identical to one another. The charge and mass of electrons are not average values; rather, they are unique values that all electrons have. This is true of other fundamental entities at the submicroscopic level. All protons are identical to one another, and so on.

The Nucleus

Here, we examine the first direct evidence of the size and mass of the nucleus. In later chapters, we will examine many other aspects of nuclear physics, but the basic information on nuclear size and mass is so important to understanding the atom that we consider it here.

Nuclear radioactivity was discovered in 1896, and it was soon the subject of intense study by a number of the best scientists in the world. Among them was New Zealander Lord Ernest Rutherford, who made numerous fundamental discoveries and earned the title of "father of nuclear physics." Born in Nelson, Rutherford did his postgraduate studies at the Cavendish Laboratories in England before taking up a position at McGill University in Canada where he did the work that earned him a Nobel Prize in Chemistry in 1908. In the area of atomic and nuclear physics, there is much overlap between chemistry and physics, with physics providing the fundamental enabling theories. He returned to England in later years and had six future Nobel Prize winners as students. Rutherford used nuclear radiation to directly examine the size and mass of the atomic nucleus. The experiment he devised is shown in [link]. A radioactive source that emits alpha radiation was placed in a lead container with a hole in one side to produce a beam of alpha particles, which are a type of ionizing radiation ejected by the nuclei of a radioactive source. A thin gold foil was placed in the beam, and the scattering of the alpha particles was observed by the glow they caused when they struck a phosphor screen.

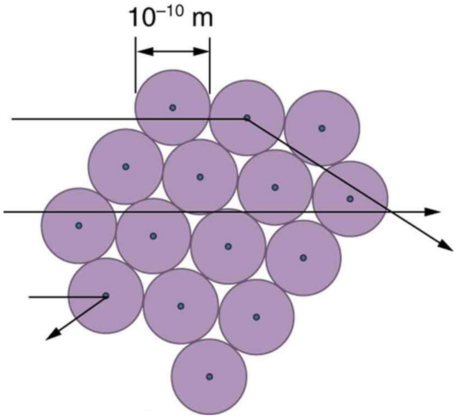


Rutherford’s experiment gave direct evidence for the size and mass of the nucleus by scattering alpha particles from a thin gold foil. Alpha particles with energies of about 5 MeV are emitted from a radioactive source (which is a small metal container in which a specific amount of a radioactive material is sealed), are collimated into a beam, and fall upon the foil. The number of particles that penetrate the foil or scatter to various angles indicates that gold nuclei are very small and contain nearly all of the gold atom’s mass. This is particularly indicated by the alpha particles that scatter to very large angles, much like a soccer ball bouncing off a goalie’s head.

Alpha particles were known to be the doubly charged positive nuclei of helium atoms that had kinetic energies on the order of 5 MeV when emitted in nuclear decay, which is the disintegration of the nucleus of an unstable nuclide by the spontaneous emission of charged particles. These particles interact with matter mostly via the Coulomb force, and the manner in which they scatter from nuclei can reveal nuclear size and mass. This is analogous to observing how a bowling ball is scattered by an object you cannot see directly. Because the alpha particle’s energy is so large compared with the typical energies associated with atoms (MeV versus eV), you would expect the alpha particles to simply crash through a thin foil much like a supersonic bowling ball would crash through a few dozen rows of bowling pins. Thomson had envisioned the atom to be a small sphere in which equal amounts of positive and negative charge were distributed evenly. The incident massive alpha particles would suffer only small deflections in such a model. Instead, Rutherford and his collaborators found that alpha particles occasionally were scattered to large angles, some even back in the direction from which they came! Detailed analysis using conservation of momentum and energy—particularly of the small number that came straight back—implied that gold nuclei are very small compared with the size of a gold atom, contain almost all of the atom’s mass, and are tightly bound. Since the gold nucleus is several times more massive than the alpha particle, a head-on collision would scatter the alpha particle straight back toward the source. In addition, the smaller the nucleus, the fewer alpha particles that would hit one head on.

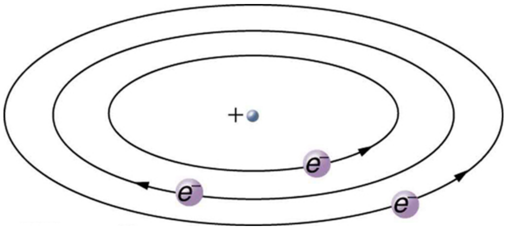
Although the results of the experiment were published by his colleagues in 1909, it took Rutherford two years to convince himself of their meaning. Like Thomson before him, Rutherford was reluctant to accept such radical results. Nature on a small scale is so unlike our classical world that even those at the forefront of discovery are sometimes surprised. Rutherford later wrote: “It was almost as incredible as if you fired a 15-inch shell at a piece of tissue paper and it came back and hit you. On consideration, I realized that this scattering backwards ... [meant] ... the greatest part of the mass of the atom was concentrated in a tiny nucleus.” In 1911, Rutherford published his analysis together with a proposed model of the atom. The size of the nucleus was determined to be about 10^{-15} m, or 100,000 times smaller than the atom. This implies a huge density, on the order of 10^{15} g/cm^3, vastly unlike any macroscopic matter. Also implied is the existence of previously unknown nuclear forces to counteract the huge repulsive Coulomb forces among the positive charges in the nucleus. Huge forces would also be consistent with the large energies emitted in nuclear radiation.

The small size of the nucleus also implies that the atom is mostly empty inside. In fact, in Rutherford’s experiment, most alphas went straight through the gold foil with very little scattering, since electrons have such small masses and since the atom was mostly empty with nothing for the alpha to hit. There were already hints of this at the time Rutherford performed his experiments, since energetic electrons had been observed to penetrate thin foils more easily than expected. [link] shows a schematic of the atoms in a thin foil with circles representing the size of the atoms (about 10^{-10} m) and dots representing the nuclei. (The dots are not to scale—if they were, you would need a microscope to see them.) Most alpha particles miss the small nuclei and are only slightly scattered by electrons. Occasionally, (about once in 8000 times in Rutherford’s experiment), an alpha hits a nucleus head-on and is scattered straight backward.



An expanded view of the atoms in the gold foil in Rutherford’s experiment. Circles represent the atoms (about $10^{-10}\ \text{m}$ in diameter), while the dots represent the nuclei (about $10^{-15}\ \text{m}$ in diameter). To be visible, the dots are much larger than scale. Most alpha particles crash through but are relatively unaffected because of their high energy and the electron’s small mass. Some, however, head straight toward a nucleus and are scattered straight back. A detailed analysis gives the size and mass of the nucleus.

Based on the size and mass of the nucleus revealed by his experiment, as well as the mass of electrons, Rutherford proposed the **planetary model of the atom**. The planetary model of the atom pictures low-mass electrons orbiting a large-mass nucleus. The sizes of the electron orbits are large compared with the size of the nucleus, with mostly vacuum inside the atom. This picture is analogous to how low-mass planets in our solar system orbit the large-mass Sun at distances large compared with the size of the sun. In the atom, the attractive Coulomb force is analogous to gravitation in the planetary system. (See [link].) Note that a model or mental picture is needed to explain experimental results, since the atom is too small to be directly observed with visible light.



Rutherford’s planetary model of the atom incorporates the characteristics of the nucleus, electrons, and the size of the atom. This model was the first to recognize the structure of atoms, in which low-mass electrons orbit a very small, massive nucleus in orbits much larger than the nucleus. The atom is mostly empty and is analogous to our planetary system.

Rutherford’s planetary model of the atom was crucial to understanding the characteristics of atoms, and their interactions and energies, as we shall see in the next few sections. Also, it was an indication of how different nature is from the familiar classical world on the small, quantum mechanical scale. The discovery of a substructure to all matter in the form of atoms and molecules was now being taken a step further to reveal a substructure of atoms that was simpler than the 92 elements then known. We have continued to search for deeper substructures, such as those inside the nucleus, with some success. In later chapters, we will follow this quest in the discussion of quarks and other elementary particles, and we will look at the direction the search seems now to be heading.

Note:
PhET Explorations: Rutherford Scattering
How did Rutherford figure out the structure of the atom without being able to see it? Simulate the famous experiment in which he disproved the Plum Pudding model of the atom by observing alpha particles bouncing off atoms and determining that they must have a small core.

https://phet.colorado.edu/sims/html/rutherford-scattering/latest/rutherford-scattering_en.html

Section Summary

- Atoms are composed of negatively charged electrons, first proved to exist in cathode-ray-tube experiments, and a positively charged nucleus.
- All electrons are identical and have a charge-to-mass ratio of
  Equation:

$$\frac{q_e}{m_e} = -1.76 \times 10^{11}\ \text{C/kg}.$$

- The positive charge in the nuclei is carried by particles called protons, which have a charge-to-mass ratio of
  Equation:

$$\frac{q_p}{m_p} = 9.57 \times 10^7 \text{ C/kg.}$$

- Mass of electron,
  Equation:

$$m_e = 9.11 \times 10^{-31} \text{ kg.}$$

- Mass of proton,
  Equation:

$$m_p = 1.67 \times 10^{-27} \text{ kg.}$$

- The planetary model of the atom pictures electrons orbiting the nucleus in the same way that planets orbit the sun.

Conceptual Questions

Exercise:

Problem:

What two pieces of evidence allowed the first calculation of m_e, the mass of the electron?

(a) The ratios q_e/m_e and q_p/m_p.

(b) The values of q_e and E_B.

(c) The ratio q_e/m_e and q_e.

Justify your response.

Exercise:

Problem:

How do the allowed orbits for electrons in atoms differ from the allowed orbits for planets around the sun? Explain how the correspondence principle applies here.

Problem Exercises

Exercise:

Problem:

Rutherford found the size of the nucleus to be about 10^{-15} m. This implied a huge density. What would this density be for gold?

Solution:

$6 \times 10^{20}\text{ kg/m}^3$

Exercise:

Problem:

In Millikan’s oil-drop experiment, one looks at a small oil drop held motionless between two plates. Take the voltage between the plates to be 2033 V, and the plate separation to be 2.00 cm. The oil drop (of density 0.81 g/cm^3) has a diameter of $4.0 \times 10^{-6}\text{ m}$. Find the charge on the drop, in terms of electron units.

Exercise:

Problem:

(a) An aspiring physicist wants to build a scale model of a hydrogen atom for her science fair project. If the atom is 1.00 m in diameter, how big should she try to make the nucleus?

(b) How easy will this be to do?

Solution:

(a) 10.0 μm

(b) It isn’t hard to make one of approximately this size. It would be harder to make it exactly 10.0 μm.

Glossary

cathode-ray tube
a vacuum tube containing a source of electrons and a screen to view images

planetary model of the atom
the most familiar model or illustration of the structure of the atom

Bohr’s Theory of the Hydrogen Atom

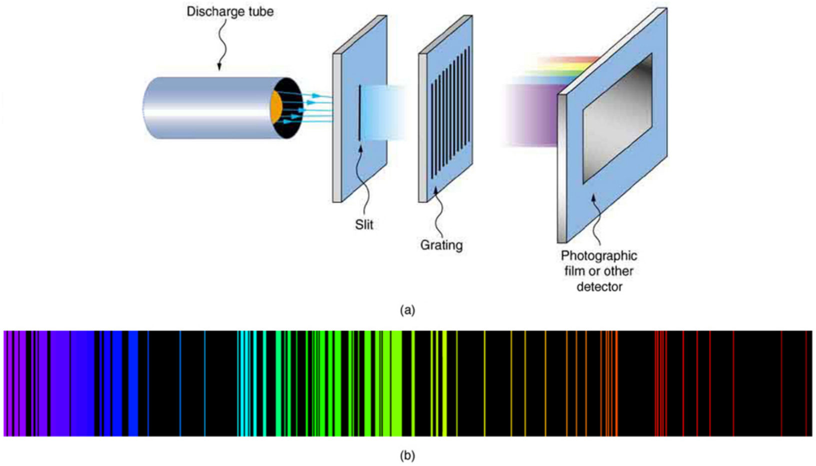
- Describe the mysteries of atomic spectra.
- Explain Bohr’s theory of the hydrogen atom.
- Explain Bohr’s planetary model of the atom.
- Illustrate energy state using the energy-level diagram.
- Describe the triumphs and limits of Bohr’s theory.

The great Danish physicist Niels Bohr (1885–1962) made immediate use of Rutherford’s planetary model of the atom. ([link]). Bohr became convinced of its validity and spent part of 1912 at Rutherford’s laboratory. In 1913, after returning to Copenhagen, he began publishing his theory of the simplest atom, hydrogen, based on the planetary model of the atom. For decades, many questions had been asked about atomic characteristics. From their sizes to their spectra, much was known about atoms, but little had been explained in terms of the laws of physics. Bohr’s theory explained the atomic spectrum of hydrogen and established new and broadly applicable principles in quantum mechanics.

Niels Bohr, Danish physicist, used the planetary model of the atom to explain the atomic spectrum and size of the hydrogen atom. His many contributions to the development of atomic physics and quantum mechanics, his personal influence on many students and colleagues, and his personal integrity, especially in the face of Nazi oppression, earned him a prominent place in history. (credit: Unknown Author, via Wikimedia Commons)

Mysteries of Atomic Spectra

As noted in Quantization of Energy , the energies of some small systems are quantized. Atomic and molecular emission and absorption spectra have been known for over a century to be discrete (or quantized). (See [link].) Maxwell and others had realized that there must be a connection between the spectrum of an atom and its structure, something like the resonant frequencies of musical instruments. But, in spite of years of efforts by many great minds, no one had a workable theory. (It was a running joke that any theory of atomic and molecular spectra could be destroyed by throwing a book of data at it, so complex were the spectra.) Following Einstein’s proposal of photons with quantized energies directly proportional to their wavelengths, it became even more evident that electrons in atoms can exist only in discrete orbits.



Part (a) shows, from left to right, a discharge tube, slit, and diffraction grating producing a line spectrum. Part (b) shows the emission line spectrum for iron. The discrete lines imply quantized energy states for the atoms that produce them. The line spectrum for each element is unique, providing a powerful and much used analytical tool, and many line spectra were well known for many years before they could be explained with physics. (credit for (b): Yttrium91, Wikimedia Commons)

In some cases, it had been possible to devise formulas that described the emission spectra. As you might expect, the simplest atom—hydrogen, with its single electron—has a relatively simple spectrum. The hydrogen spectrum had been observed in the infrared (IR), visible, and ultraviolet (UV), and several series of spectral lines had been observed. (See [link].) These series are named after early researchers who studied them in particular depth.

The observed **hydrogen-spectrum wavelengths** can be calculated using the following formula:

Equation:

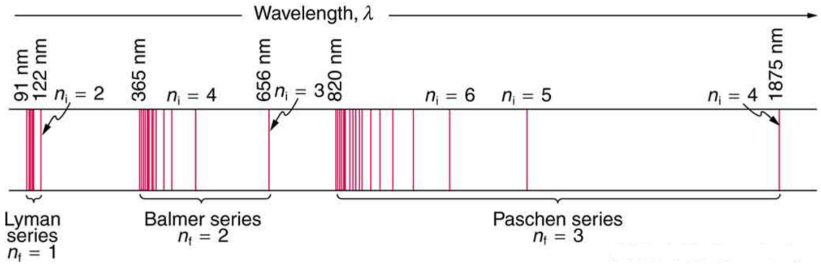
$$\frac{1}{\lambda} = R\left(\frac{1}{n_\text{f}^2} - \frac{1}{n_\text{i}^2}\right),$$

where λ is the wavelength of the emitted EM radiation and R is the **Rydberg constant**, determined by the experiment to be

Equation:

$$R = 1.097 \times 10^7/\text{m (or m}^{-1}).$$

The constant n_f is a positive integer associated with a specific series. For the Lyman series, $n_\text{f} = 1$; for the Balmer series, $n_\text{f} = 2$; for the Paschen series, $n_\text{f} = 3$; and so on. The Lyman series is entirely in the UV, while part of the Balmer series is visible with the remainder UV. The Paschen series and all the rest are entirely IR. There are apparently an unlimited number of series, although they lie progressively farther into the infrared and become difficult to observe as n_f increases. The constant n_i is a positive integer, but it must be greater than n_f. Thus, for the Balmer series, $n_\text{f} = 2$ and $n_\text{i} = 3, 4, 5, 6, ...$. Note that n_i can approach infinity. While the formula in the wavelengths equation was just a recipe designed to fit data and was not based on physical principles, it did imply a deeper meaning. Balmer first devised the formula for his series alone, and it was later found to describe all the other series by using different values of n_f. Bohr was the first to comprehend the deeper meaning. Again, we see the interplay between experiment and theory in physics. Experimentally, the spectra were well established, an equation was found to fit the experimental data, but the theoretical foundation was missing.



A schematic of the hydrogen spectrum shows several series named for those who contributed most to their determination. Part of the Balmer series is in the visible spectrum, while the Lyman series is entirely in the UV, and the Paschen series and others are in the IR. Values of n_f and n_i are shown for some of the lines.

Example:
Calculating Wave Interference of a Hydrogen Line

What is the distance between the slits of a grating that produces a first-order maximum for the second Balmer line at an angle of 15º?

Strategy and Concept

For an Integrated Concept problem, we must first identify the physical principles involved. In this example, we need to know (a) the wavelength of light as well as (b) conditions for an interference maximum for the pattern from a double slit. Part (a) deals with a topic of the present chapter, while part (b) considers the wave interference material of Wave Optics.

Solution for (a)

Hydrogen spectrum wavelength. The Balmer series requires that $n_{\text{f}} = 2$. The first line in the series is taken to be for $n_{\text{i}} = 3$, and so the second would have $n_{\text{i}} = 4$.

The calculation is a straightforward application of the wavelength equation. Entering the determined values for n_{f} and n_{i} yields

Equation:

$$\begin{aligned} \frac{1}{\lambda} &= R\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right) \\ &= \left(1.097 \times 10^7 \text{ m}^{-1}\right)\left(\frac{1}{2^2} - \frac{1}{4^2}\right) \\ &= 2.057 \times 10^6 \text{ m}^{-1}. \end{aligned}$$

Inverting to find λ gives

Equation:

$$\begin{aligned} \lambda &= \frac{1}{2.057 \times 10^6 \text{ m}^{-1}} = 486 \times 10^{-9} \text{ m} \\ &= 486 \text{ nm}. \end{aligned}$$

Discussion for (a)

This is indeed the experimentally observed wavelength, corresponding to the second (blue-green) line in the Balmer series. More impressive is the fact that the same simple recipe predicts *all* of the hydrogen spectrum lines, including new ones observed in subsequent experiments. What is nature telling us?

Solution for (b)

Double-slit interference (Wave Optics). To obtain constructive interference for a double slit, the path length difference from two slits must be an integral multiple of the wavelength. This condition was expressed by the equation

Equation:

$$d \sin \theta = m\lambda,$$

where d is the distance between slits and θ is the angle from the original direction of the beam. The number m is the order of the interference; $m = 1$ in this example. Solving for d and entering known values yields

Equation:

$$d = \frac{(1)(486\text{ nm})}{\sin 15^{\text{o}}} = 1.88 \times 10^{-6}\text{ m.}$$

Discussion for (b)

This number is similar to those used in the interference examples of Introduction to Quantum Physics (and is close to the spacing between slits in commonly used diffraction glasses).

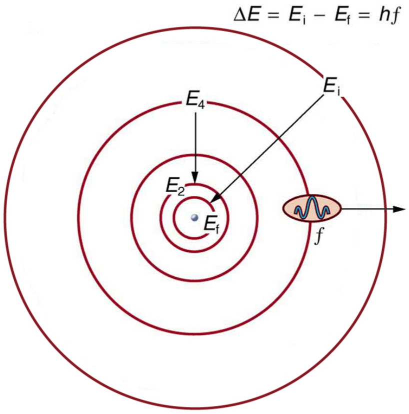
Bohr’s Solution for Hydrogen

Bohr was able to derive the formula for the hydrogen spectrum using basic physics, the planetary model of the atom, and some very important new proposals. His first proposal is that only certain orbits are allowed: we say that *the orbits of electrons in atoms are quantized*. Each orbit has a different energy, and electrons can move to a higher orbit by absorbing energy and drop to a lower orbit by emitting energy. If the orbits are quantized, the amount of energy absorbed or emitted is also quantized, producing discrete spectra. Photon absorption and emission are among the primary methods of transferring energy into and out of atoms. The energies of the photons are quantized, and their energy is explained as being equal to the change in energy of the electron when it moves from one orbit to another. In equation form, this is

Equation:

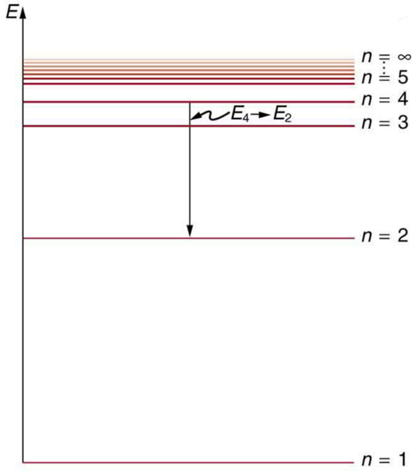
$$\Delta E = hf = E_{\text{i}} - E_{\text{f}}.$$

Here, ΔE is the change in energy between the initial and final orbits, and hf is the energy of the absorbed or emitted photon. It is quite logical (that is, expected from our everyday experience) that energy is involved in changing orbits. A blast of energy is required for the space shuttle, for example, to climb to a higher orbit. What is not expected is that atomic orbits should be quantized. This is not observed for satellites or planets, which can have any orbit given the proper energy. (See [link].)



The planetary model of the atom, as modified by Bohr, has the orbits of the electrons quantized. Only certain orbits are allowed, explaining why atomic spectra are discrete (quantized). The energy carried away from an atom by a photon comes from the electron dropping from one allowed orbit to another and is thus quantized. This is likewise true for atomic absorption of photons.

[link] shows an **energy-level diagram**, a convenient way to display energy states. In the present discussion, we take these to be the allowed energy levels of the electron. Energy is plotted vertically with the lowest or ground state at the bottom and with excited states above. Given the energies of the lines in an atomic spectrum, it is possible (although sometimes very difficult) to determine the energy levels of an atom. Energy-level diagrams are used for many systems, including molecules and nuclei. A theory of the atom or any other system must predict its energies based on the physics of the system.



An energy-level diagram plots energy vertically and is useful in visualizing the energy states of a system and the transitions between them. This diagram is for the hydrogen-atom electrons, showing a transition between two orbits having energies E_4 and E_2.

Bohr was clever enough to find a way to calculate the electron orbital energies in hydrogen. This was an important first step that has been improved upon, but it is well worth repeating here, because it does correctly describe many characteristics of hydrogen. Assuming circular orbits, Bohr proposed that the **angular momentum L of an electron in its orbit is quantized**, that is, it has only specific, discrete values. The value for L is given by the formula

Equation:

$$L = m_e \mathrm{vr}_n = n\frac{h}{2\pi}(n = 1, 2, 3, \ldots),$$

where L is the angular momentum, m_e is the electron's mass, r_n is the radius of the n th orbit, and h is Planck's constant. Note that angular momentum is $L = I\omega$. For a small object at a radius r, $I = mr^2$ and $\omega = v/r$, so that $L = \left(mr^2\right)(v/r) = mvr$. Quantization says that this value of mvr can only be equal to $h/2$, $2h/2$, $3h/2$, etc. At the time, Bohr himself did not know why angular momentum should be quantized, but using this assumption he was able to calculate the energies in the hydrogen spectrum, something no one else had done at the time.

From Bohr’s assumptions, we will now derive a number of important properties of the hydrogen atom from the classical physics we have covered in the text. We start by noting the centripetal force causing the electron to follow a circular path is supplied by the Coulomb force. To be more general, we note that this analysis is valid for any single-electron atom. So, if a nucleus has Z protons ($Z = 1$ for hydrogen, 2 for helium, etc.) and only one electron, that atom is called a **hydrogen-like atom**. The spectra of hydrogen-like ions are similar to hydrogen, but shifted to higher energy by the greater attractive force between the electron and nucleus. The magnitude of the centripetal force is $m_e v^2/r_n$, while the Coulomb force is $k(\mathrm{Zq}_e)(q_e)/r_n^2$. The tacit assumption here is that the nucleus is more massive than the stationary electron, and the electron orbits about it. This is consistent with the planetary model of the atom. Equating these,

Equation:

$$k\frac{Zq_e^2}{r_n^2} = \frac{m_e v^2}{r_n} \text{ (Coulomb = centripetal).}$$

Angular momentum quantization is stated in an earlier equation. We solve that equation for v, substitute it into the above, and rearrange the expression to obtain the radius of the orbit. This yields:

Equation:

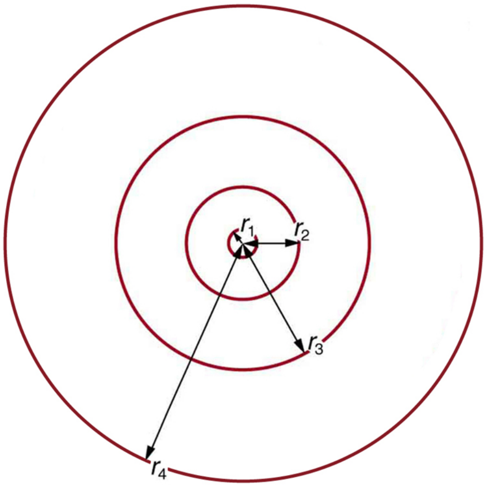
$$r_n = \frac{n^2}{Z}a_{\mathrm{B}}, \text{ for allowed orbits}(n = 1,2,3,\ldots),$$

where a_{B} is defined to be the **Bohr radius**, since for the lowest orbit $(n = 1)$ and for hydrogen $(Z = 1)$, $r_1 = a_{\mathrm{B}}$. It is left for this chapter’s Problems and Exercises to show that the Bohr radius is

Equation:

$$a_{\mathrm{B}} = \frac{h^2}{4\pi^2 m_e k q_e^2} = 0.529 \times 10^{-10}\text{ m.}$$

These last two equations can be used to calculate the **radii of the allowed (quantized) electron orbits in any hydrogen-like atom**. It is impressive that the formula gives the correct size of hydrogen, which is measured experimentally to be very close to the Bohr radius. The earlier equation also tells us that the orbital radius is proportional to n^2, as illustrated in [link].



The allowed electron orbits in hydrogen have the radii shown. These radii were first calculated by Bohr and are given by the equation $r_n = \frac{n^2}{Z} a_{\mathrm{B}}$. The lowest orbit has the experimentally verified diameter of a hydrogen atom.

To get the electron orbital energies, we start by noting that the electron energy is the sum of its kinetic and potential energy:

Equation:

$$E_n = \mathrm{KE} + \mathrm{PE}.$$

Kinetic energy is the familiar $\mathrm{KE} = (1/2) m_e v^2$, assuming the electron is not moving at relativistic speeds. Potential energy for the electron is electrical, or $\mathrm{PE} = q_e V$, where V is the potential due to the nucleus, which looks like a point charge. The nucleus has a positive charge Zq_e ; thus, $V = \mathrm{kZq}_e / r_n$, recalling an earlier equation for the potential due to a point charge. Since the electron’s charge is negative, we see that $\mathrm{PE} = -\mathrm{kZq}_e / r_n$. Entering the expressions for KE and PE, we find

Equation:

$$E_n = \frac{1}{2} m_e v^2 - k \frac{Z q_e^2}{r_n}.$$

Now we substitute r_n and v from earlier equations into the above expression for energy. Algebraic manipulation yields

Equation:

$$E_n = -\frac{Z^2}{n^2}E_0(n = 1, 2, 3, ...)$$

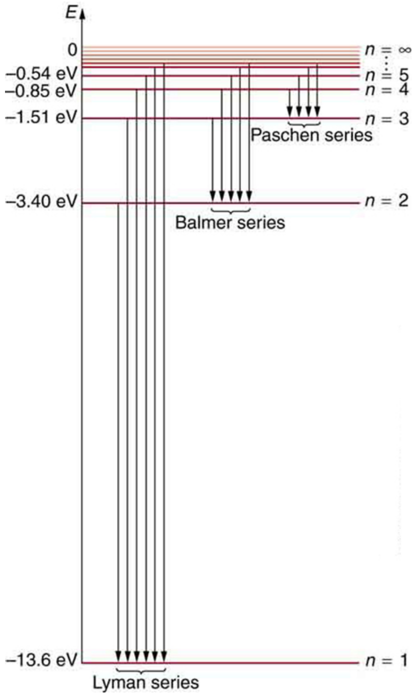
for the orbital **energies of hydrogen-like atoms**. Here, E_0 is the **ground-state energy** $(n = 1)$ for hydrogen $(Z = 1)$ and is given by
Equation:

$$E_0 = \frac{2\pi^2 q_e^4 m_e k^2}{h^2} = 13.6\text{ eV}.$$

Thus, for hydrogen,
Equation:

$$E_n = -\frac{13.6\text{ eV}}{n^2}(n = 1, 2, 3, ...).$$

[link] shows an energy-level diagram for hydrogen that also illustrates how the various spectral series for hydrogen are related to transitions between energy levels.



Energy-level diagram for

hydrogen showing the Lyman, Balmer, and Paschen series of transitions. The orbital energies are calculated using the above equation, first derived by Bohr.

Electron total energies are negative, since the electron is bound to the nucleus, analogous to being in a hole without enough kinetic energy to escape. As n approaches infinity, the total energy becomes zero. This corresponds to a free electron with no kinetic energy, since r_n gets very large for large n, and the electric potential energy thus becomes zero. Thus, 13.6 eV is needed to ionize hydrogen (to go from –13.6 eV to 0, or unbound), an experimentally verified number. Given more energy, the electron becomes unbound with some kinetic energy. For example, giving 15.0 eV to an electron in the ground state of hydrogen strips it from the atom and leaves it with 1.4 eV of kinetic energy.

Finally, let us consider the energy of a photon emitted in a downward transition, given by the equation to be
Equation:

$$\Delta E = \text{hf} = E_{\text{i}} - E_{\text{f}}.$$

Substituting $E_n = (-13.6\text{ eV}/n^2)$, we see that
Equation:

$$\text{hf} = (13.6\text{ eV})\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right).$$

Dividing both sides of this equation by hc gives an expression for $1/\lambda$:
Equation:

$$\frac{\text{hf}}{\text{hc}} = \frac{f}{c} = \frac{1}{\lambda} = \frac{(13.6\text{ eV})}{hc}\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right).$$

It can be shown that
Equation:

$$\left(\frac{13.6\text{ eV}}{hc}\right) = \frac{(13.6\text{ eV})\left(1.602\times10^{-19}\text{ J/eV}\right)}{\left(6.626\times10^{-34}\text{ J·s}\right)\left(2.998\times10^{8}\text{ m/s}\right)} = 1.097\times10^{7}\text{ m}^{-1} = R$$

is the **Rydberg constant**. Thus, we have used Bohr's assumptions to derive the formula first proposed by Balmer years earlier as a recipe to fit experimental data.

Equation:

$$\frac{1}{\lambda} = R\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right)$$

We see that Bohr's theory of the hydrogen atom answers the question as to why this previously known formula describes the hydrogen spectrum. It is because the energy levels are proportional to $1/n^2$, where n is a non-negative integer. A downward transition releases energy, and so n_{i} must be greater than n_{f}. The various series are those where the transitions end on a certain level. For the Lyman series, $n_{\text{f}} = 1$ — that is, all the transitions end in the ground state (see also [link]). For the Balmer series, $n_{\text{f}} = 2$, or all the transitions end in the first excited state; and so on. What was once a recipe is now based in physics, and something new is emerging—angular momentum is quantized.

Triumphs and Limits of the Bohr Theory

Bohr did what no one had been able to do before. Not only did he explain the spectrum of hydrogen, he correctly calculated the size of the atom from basic physics. Some of his ideas are broadly applicable. Electron orbital energies are quantized in all atoms and molecules. Angular momentum is quantized. The electrons do not spiral into the nucleus, as expected classically (accelerated charges radiate, so that the electron orbits classically would decay quickly, and the electrons would sit on the nucleus—matter would collapse). These are major triumphs.

But there are limits to Bohr's theory. It cannot be applied to multielectron atoms, even one as simple as a two-electron helium atom. Bohr's model is what we call *semiclassical*. The orbits are quantized (nonclassical) but are assumed to be simple circular paths (classical). As quantum mechanics was developed, it became clear that there are no well-defined orbits; rather, there are clouds of probability. Bohr's theory also did not explain that some spectral lines are doublets (split into two) when examined closely. We shall examine many of these aspects of quantum mechanics in more detail, but it should be kept in mind that Bohr did not fail. Rather, he made very important steps along the path to greater knowledge and laid the foundation for all of atomic physics that has since evolved.

Note:

PhET Explorations: Models of the Hydrogen Atom

How did scientists figure out the structure of atoms without looking at them? Try out different models by shooting light at the atom. Check how the prediction of the model matches the experimental results.

https://archive.cnx.org/specials/d77cc1d0-33e4-11e6-b016-6726afecd2be/hydrogen-atom/#sim-hydrogen-atom

Section Summary

- The planetary model of the atom pictures electrons orbiting the nucleus in the way that planets orbit the sun. Bohr used the planetary model to develop the first reasonable theory of hydrogen, the simplest atom. Atomic and molecular spectra are quantized, with hydrogen spectrum wavelengths given by the formula
  Equation:
  $$\frac{1}{\lambda} = R\left(\frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}\right),$$
  where λ is the wavelength of the emitted EM radiation and R is the Rydberg constant, which has the value
  Equation:
  $$R = 1.097 \times 10^7 \text{ m}^{-1}.$$
- The constants n_{i} and n_{f} are positive integers, and n_{i} must be greater than n_{f}.
- Bohr correctly proposed that the energy and radii of the orbits of electrons in atoms are quantized, with energy for transitions between orbits given by
  Equation:
  $$\Delta E = hf = E_{\text{i}} - E_{\text{f}},$$
  where ΔE is the change in energy between the initial and final orbits and hf is the energy of an absorbed or emitted photon. It is useful to plot orbital energies on a vertical graph called an energy-level diagram.
- Bohr proposed that the allowed orbits are circular and must have quantized orbital angular momentum given by
  Equation:
  $$L = m_e v r_n = n\frac{h}{2\pi}(n = 1, 2, 3 \ldots),$$
  where L is the angular momentum, r_n is the radius of the $n\text{th}$ orbit, and h is Planck's constant. For all one-electron (hydrogen-like) atoms, the radius of an orbit is given by
  Equation:
  $$r_n = \frac{n^2}{Z}a_{\text{B}}(\text{allowed orbits } n = 1, 2, 3, \ldots),$$

Z is the atomic number of an element (the number of electrons is has when neutral) and a_{B} is defined to be the Bohr radius, which is
Equation:

$$a_{\text{B}} = \frac{h^2}{4\pi^2 m_e k q_e^2} = 0.529 \times 10^{-10}\ \text{m}.$$

- Furthermore, the energies of hydrogen-like atoms are given by
  Equation:

  $$E_n = -\frac{Z^2}{n^2} E_0 (n = 1, 2, 3\ ...),$$

  where E_0 is the ground-state energy and is given by
  Equation:

  $$E_0 = \frac{2\pi^2 q_e^4 m_e k^2}{h^2} = 13.6\ \text{eV}.$$

  Thus, for hydrogen,
  Equation:

  $$E_n = -\frac{13.6\ \text{eV}}{n^2} (n, =, 1, 2, 3\ ...).$$

- The Bohr Theory gives accurate values for the energy levels in hydrogen-like atoms, but it has been improved upon in several respects.

Conceptual Questions

Exercise:

Problem:

How do the allowed orbits for electrons in atoms differ from the allowed orbits for planets around the sun? Explain how the correspondence principle applies here.

Exercise:

Problem:

Explain how Bohr’s rule for the quantization of electron orbital angular momentum differs from the actual rule.

Exercise:

Problem:

What is a hydrogen-like atom, and how are the energies and radii of its electron orbits related to those in hydrogen?

Problems & Exercises

Exercise:

Problem:

By calculating its wavelength, show that the first line in the Lyman series is UV radiation.

Solution:

$\frac{1}{\lambda} = R\left(\frac{1}{n_\mathrm{f}^2} - \frac{1}{n_\mathrm{i}^2}\right) \Rightarrow \lambda = \frac{1}{R}\left[\frac{(n_\mathrm{i} \cdot n_\mathrm{f})^2}{n_\mathrm{i}^2 - n_\mathrm{f}^2}\right]; n_\mathrm{i} = 2, n_\mathrm{f} = 1,$ so that

$\lambda = \left(\frac{\mathrm{m}}{1.097 \times 10^7}\right)\left[\frac{(2 \times 1)^2}{2^2 - 1^2}\right] = 1.22 \times 10^{-7}\ \mathrm{m} = 122\ \mathrm{nm}$, which is UV radiation.

Exercise:

Problem:

Find the wavelength of the third line in the Lyman series, and identify the type of EM radiation.

Exercise:

Problem:

Look up the values of the quantities in $a_\mathrm{B} = \frac{h^2}{4\pi^2 m_e k q_e^2}$, and verify that the Bohr radius a_B is 0.529×10^{-10} m.

Solution:

$$a_\mathrm{B} = \frac{h^2}{4\pi^2 m_e k Z q_e^2} = \frac{(6.626 \times 10^{-34}\ \mathrm{J \cdot s})^2}{4\pi^2 (9.109 \times 10^{-31}\ \mathrm{kg})(8.988 \times 10^9\ \mathrm{N \cdot m^2/C^2})(1)(1.602 \times 10^{-19}\ \mathrm{C})^2} = 0.529 \times 10^{-10}\ \mathrm{m}$$

Exercise:

Problem: Verify that the ground state energy E_0 is 13.6 eV by using $E_0 = \frac{2\pi^2 q_e^4 m_e k^2}{h^2}$.

Exercise:

Problem:

If a hydrogen atom has its electron in the $n = 4$ state, how much energy in eV is needed to ionize it?

Solution:

0.850 eV

Exercise:

Problem:

A hydrogen atom in an excited state can be ionized with less energy than when it is in its ground state. What is n for a hydrogen atom if 0.850 eV of energy can ionize it?

Exercise:

Problem:

Find the radius of a hydrogen atom in the $n = 2$ state according to Bohr's theory.

Solution:

2.12×10^{-10} m

Exercise:

Problem:

Show that $(13.6\text{ eV})/\text{hc} = 1.097 \times 10^{7}\text{ m} = R$ (Rydberg's constant), as discussed in the text.

Exercise:

Problem:

What is the smallest-wavelength line in the Balmer series? Is it in the visible part of the spectrum?

Solution:

365 nm

It is in the ultraviolet.

Exercise:

Problem:

Show that the entire Paschen series is in the infrared part of the spectrum. To do this, you only need to calculate the shortest wavelength in the series.

Exercise:

Problem:

Do the Balmer and Lyman series overlap? To answer this, calculate the shortest-wavelength Balmer line and the longest-wavelength Lyman line.

Solution:

No overlap

365 nm

122 nm

Exercise:

Problem:

(a) Which line in the Balmer series is the first one in the UV part of the spectrum?

(b) How many Balmer series lines are in the visible part of the spectrum?

(c) How many are in the UV?

Exercise:

Problem:

A wavelength of $4.653\ \mu\text{m}$ is observed in a hydrogen spectrum for a transition that ends in the $n_{\text{f}} = 5$ level. What was n_{i} for the initial level of the electron?

Solution:

7

Exercise:

Problem:

A singly ionized helium ion has only one electron and is denoted He^{+}. What is the ion's radius in the ground state compared to the Bohr radius of hydrogen atom?

Exercise:

Problem:

A beryllium ion with a single electron (denoted Be^{3+}) is in an excited state with radius the same as that of the ground state of hydrogen.

(a) What is n for the Be^{3+} ion?

(b) How much energy in eV is needed to ionize the ion from this excited state?

Solution:

(a) 2

(b) 54.4 eV

Exercise:

Problem:

Atoms can be ionized by thermal collisions, such as at the high temperatures found in the solar corona. One such ion is C^{+5}, a carbon atom with only a single electron.

(a) By what factor are the energies of its hydrogen-like levels greater than those of hydrogen?

(b) What is the wavelength of the first line in this ion’s Paschen series?

(c) What type of EM radiation is this?

Exercise:

Problem:

Verify Equations $r_n = \frac{n^2}{Z}a_{\mathrm{B}}$ and $a_{\mathrm{B}} = \frac{h^2}{4\pi^2 m_e k q_e^2} = 0.529 \times 10^{-10}\ \mathrm{m}$ using the approach stated in the text. That is, equate the Coulomb and centripetal forces and then insert an expression for velocity from the condition for angular momentum quantization.

Solution:

$\frac{kZq_e^2}{r_n^2} = \frac{m_eV^2}{r_n}$, so that $r_n = \frac{kZq_e^2}{m_eV^2} = \frac{kZq_e^2}{m_e}\frac{1}{V^2}$. From the equation $m_evr_n = n\frac{h}{2\pi}$, we can substitute for the velocity, giving: $r_n = \frac{kZq_e^2}{m_e}\cdot\frac{4\pi^2m_e^2r_n^2}{n^2h^2}$ so that $r_n = \frac{n^2}{Z}\frac{h^2}{4\pi^2m_ekq_e^2} = \frac{n^2}{Z}a_{\mathrm{B}}$, where $a_{\mathrm{B}} = \frac{h^2}{4\pi^2m_ekq_e^2}$.

Exercise:

Problem:

The wavelength of the four Balmer series lines for hydrogen are found to be 410.3, 434.2, 486.3, and 656.5 nm. What average percentage difference is found between these wavelength numbers and those predicted by $\frac{1}{\lambda} = R\left(\frac{1}{n_{\mathrm{f}}^2} - \frac{1}{n_{\mathrm{i}}^2}\right)$? It is amazing how well a simple formula (disconnected originally from theory) could duplicate this phenomenon.

Glossary

hydrogen spectrum wavelengths

the wavelengths of visible light from hydrogen; can be calculated by $\frac{1}{\lambda} = R\left(\frac{1}{n_{\mathrm{f}}^2} - \frac{1}{n_{\mathrm{i}}^2}\right)$

Rydberg constant

a physical constant related to the atomic spectra with an established value of $1.097 \times 10^7 \ \mathrm{m}^{-1}$

double-slit interference

an experiment in which waves or particles from a single source impinge upon two slits so that the resulting interference pattern may be observed

energy-level diagram

a diagram used to analyze the energy level of electrons in the orbits of an atom

Bohr radius

the mean radius of the orbit of an electron around the nucleus of a hydrogen atom in its ground state

hydrogen-like atom

any atom with only a single electron

energies of hydrogen-like atoms

Bohr formula for energies of electron states in hydrogen-like atoms: $E_n = -\frac{Z^2}{n^2}E_0(n = 1, 2, 3, \ \ldots)$

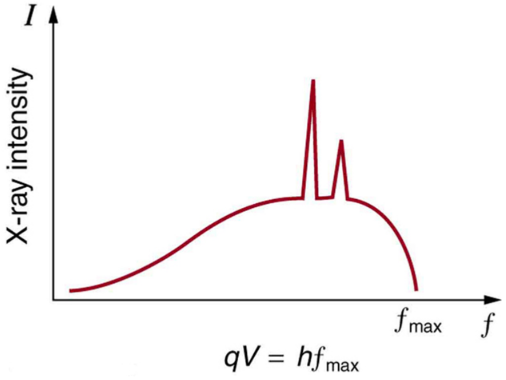
X Rays: Atomic Origins and Applications

- Define x-ray tube and its spectrum.
- Show the x-ray characteristic energy.
- Specify the use of x rays in medical observations.
- Explain the use of x rays in CT scanners in diagnostics.

Each type of atom (or element) has its own characteristic electromagnetic spectrum. **X rays** lie at the high-frequency end of an atom’s spectrum and are characteristic of the atom as well. In this section, we explore characteristic x rays and some of their important applications.

We have previously discussed x rays as a part of the electromagnetic spectrum in Photon Energies and the Electromagnetic Spectrum. That module illustrated how an x-ray tube (a specialized CRT) produces x rays. Electrons emitted from a hot filament are accelerated with a high voltage, gaining significant kinetic energy and striking the anode.

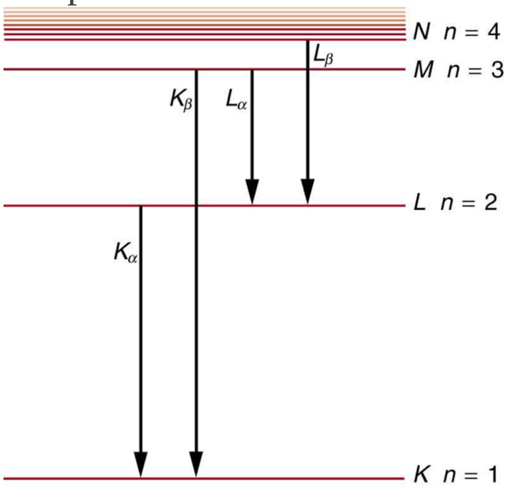
There are two processes by which x rays are produced in the anode of an x-ray tube. In one process, the deceleration of electrons produces x rays, and these x rays are called *bremsstrahlung*, or braking radiation. The second process is atomic in nature and produces *characteristic x rays*, so called because they are characteristic of the anode material. The x-ray spectrum in [link] is typical of what is produced by an x-ray tube, showing a broad curve of bremsstrahlung radiation with characteristic x-ray peaks on it.



X-ray spectrum obtained when energetic electrons strike a material, such as in the anode of a CRT. The smooth part of the spectrum is bremsstrahlung radiation, while the peaks are characteristic of the anode material. A different anode material would have characteristic x-ray peaks at different frequencies.

The spectrum in [link] is collected over a period of time in which many electrons strike the anode, with a variety of possible outcomes for each hit. The broad range of x-ray energies in the bremsstrahlung radiation indicates that an incident electron’s energy is not usually converted entirely into photon energy. The highest-energy x ray produced is one for which all of the electron’s energy was converted to photon energy. Thus the accelerating voltage and the maximum x-ray energy are related by conservation of energy. Electric potential energy is converted to kinetic energy and then to photon energy, so that $E_{\text{max}} = hf_{\text{max}} = q_e V$. Units of electron volts are convenient. For example, a 100-kV accelerating voltage produces x-ray photons with a maximum energy of 100 keV.

Some electrons excite atoms in the anode. Part of the energy that they deposit by collision with an atom results in one or more of the atom’s inner electrons being knocked into a higher orbit or the atom being ionized. When the anode’s atoms de-excite, they emit characteristic electromagnetic radiation. The most energetic of these are produced when an inner-shell vacancy is filled—that is, when an $n = 1$ or $n = 2$ shell electron has been excited to a higher level, and another electron falls into the vacant spot. A *characteristic x ray* (see Photon Energies and the Electromagnetic Spectrum) is electromagnetic (EM) radiation emitted by an atom when an inner-shell vacancy is filled. [link] shows a representative energy-level diagram that illustrates the labeling of characteristic x rays. X rays created when an electron falls into an $n = 1$ shell vacancy are called K_{α} when they come from the next higher level; that is, an $n = 2$ to $n = 1$ transition. The labels $K, L, M, \ldots$ come from the older alphabetical labeling of shells starting with K rather than using the principal quantum numbers 1, 2, 3, …. A more energetic K_{β} x ray is produced when an electron falls into an $n = 1$ shell vacancy from the $n = 3$ shell; that is, an $n = 3$ to $n = 1$ transition. Similarly, when an electron falls into the $n = 2$ shell from the $n = 3$ shell, an L_{α} x ray is created. The energies of these x rays depend on the energies of electron states in the particular atom and, thus, are characteristic of that element: every element has it own set of x-ray energies. This property can be used to identify elements, for example, to find trace (small) amounts of an element in an environmental or biological sample.



A characteristic x ray is emitted when an electron fills an inner-shell vacancy, as shown for several transitions in this approximate energy level diagram for a multiple-electron atom. Characteristic x rays are labeled according to the shell that had the vacancy and the shell from which the electron came. A K_α x ray, for example, is produced when an electron coming from the $n = 2$ shell fills the $n = 1$ shell vacancy.

Example:
Characteristic X-Ray Energy

Calculate the approximate energy of a K_α x ray from a tungsten anode in an x-ray tube.

Strategy

How do we calculate energies in a multiple-electron atom? In the case of characteristic x rays, the following approximate calculation is reasonable. Characteristic x rays are produced when an inner-shell vacancy is filled. Inner-shell electrons are nearer the nucleus than others in an atom and thus feel little net effect from the others. This is similar to what happens inside a charged conductor, where its excess charge is distributed over the surface so that it produces no electric field inside. It is reasonable to assume the inner-shell electrons have hydrogen-like energies, as given by $E_n = -\frac{Z^2}{n^2}E_0(n = 1, 2, 3, ...)$. As noted, a K_α x ray is produced by an $n = 2$ to $n = 1$ transition. Since there are two electrons in a filled K shell, a vacancy would leave one electron, so that the effective charge would be $Z - 1$ rather than Z. For tungsten, $Z = 74$, so that the effective charge is 73.

Solution

$E_n = -\frac{Z^2}{n^2}E_0(n = 1, 2, 3, ...)$ gives the orbital energies for hydrogen-like atoms to be $E_n = -(Z^2/n^2)E_{0,}$ where $E_0 = 13.6$ eV. As noted, the effective Z is 73. Now the K_α x-ray energy is given by

Equation:

$$E_{K_\alpha} = \Delta E = E_\text{i} - E_\text{f} = E_2 - E_1,$$

where

Equation:

$$E_1 = -\frac{Z^2}{1^2}E_0 = -\frac{73^2}{1}\left(13.6\text{ eV}\right) = -72.5\text{ keV}$$

and

Equation:

$$E_2 = -\frac{Z^2}{2^2}E_0 = -\frac{73^2}{4}\left(13.6\text{ eV}\right) = -18.1\text{ keV}.$$

Thus,

Equation:

$$E_{K_\alpha} = -18.1\text{ keV} - (-72.5\text{ keV}) = 54.4\text{ keV}.$$

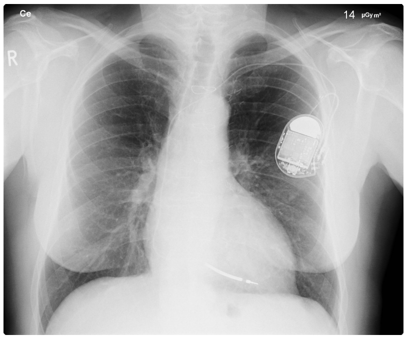
Discussion

This large photon energy is typical of characteristic x rays from heavy elements. It is large compared with other atomic emissions because it is produced when an inner-shell vacancy is filled, and inner-shell electrons are tightly bound. Characteristic x ray energies become progressively larger for heavier elements because their energy increases approximately as Z^2. Significant accelerating voltage is needed to create these inner-shell vacancies. In the case of tungsten, at least 72.5 kV is needed, because other shells are filled and you cannot simply bump one electron to a higher filled shell. Tungsten is a common anode material in x-ray tubes; so much of the energy of the impinging electrons is absorbed, raising its temperature, that a high-melting-point material like tungsten is required.

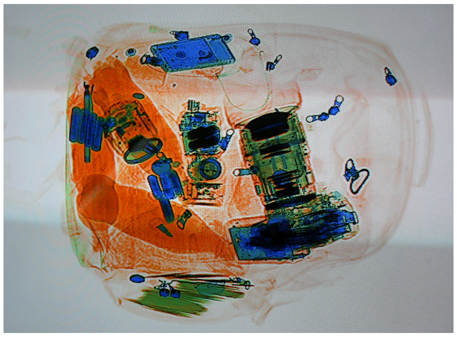
Medical and Other Diagnostic Uses of X-rays

All of us can identify diagnostic uses of x-ray photons. Among these are the universal dental and medical x rays that have become an essential part of medical diagnostics. (See [link] and [link].) X rays are also used to inspect our luggage at airports, as shown in [link], and for early detection of cracks in crucial aircraft components. An x ray is not only a noun meaning high-energy photon, it also is an image produced by x rays, and it has been made into a familiar verb—to be x-rayed.

An x-ray image reveals fillings in a person’s teeth. (credit: Dmitry G, Wikimedia Commons)



This x-ray image of a person’s chest shows many details, including an artificial pacemaker. (credit: Sunzi99, Wikimedia Commons)



This x-ray image shows the contents of a piece of luggage. The denser the material, the darker the shadow. (credit: IDuke, Wikimedia Commons)

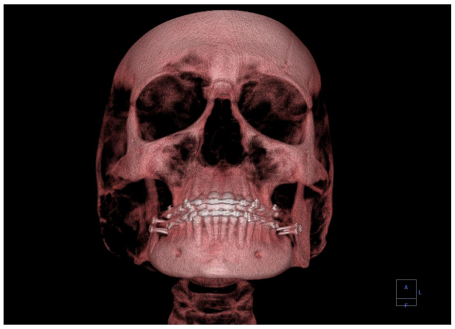
The most common x-ray images are simple shadows. Since x-ray photons have high energies, they penetrate materials that are opaque to visible light. The more energy an x-ray photon has, the more material it will penetrate. So an x-ray tube may be operated at 50.0 kV for a chest x ray, whereas it may need to be operated at 100 kV to examine a broken leg in a cast. The depth of penetration is related to the density of the material as well as to the energy of the photon. The denser the material, the fewer x-ray photons get through and the darker the shadow. Thus x rays excel at detecting breaks in bones and in imaging other physiological structures, such as some tumors, that differ in density from surrounding material. Because of their high photon energy, x rays produce significant ionization in materials and damage cells in biological organisms. Modern uses minimize exposure to the patient and eliminate exposure to others. Biological effects of x rays will be explored in the next chapter along with other types of ionizing radiation such as those produced by nuclei.

As the x-ray energy increases, the Compton effect (see Photon Momentum) becomes more important in the attenuation of the x rays. Here, the x ray scatters from an outer electron shell of the atom, giving the ejected electron some kinetic energy while losing energy itself. The probability for attenuation of the x rays depends upon the number of electrons present (the material's density) as well as the thickness of the material. Chemical composition of the medium, as characterized by its atomic number Z, is not important here. Low-energy x rays provide better contrast (sharper images). However, due to greater attenuation and less scattering, they are more absorbed by thicker materials. Greater contrast can be achieved by injecting a substance with a large atomic number, such as barium or iodine. The structure of the part of the body that contains the substance (e.g., the gastro-intestinal tract or the abdomen) can easily be seen this way.

Breast cancer is the second-leading cause of death among women worldwide. Early detection can be very effective, hence the importance of x-ray diagnostics. A mammogram cannot diagnose a malignant tumor, only give evidence of a lump or region of increased density within the breast. X-ray absorption by different types of soft tissue is very similar, so contrast is difficult; this is especially true for younger women, who typically have denser breasts. For older women who are at greater risk of developing breast cancer, the presence of more fat in the breast gives the lump or tumor more contrast. MRI (Magnetic resonance imaging) has recently been used as a supplement to conventional x rays to improve detection and eliminate false positives. The subject's radiation dose from x rays will be treated in a later chapter.

A standard x ray gives only a two-dimensional view of the object. Dense bones might hide images of soft tissue or organs. If you took another x ray from the side of the person (the first one being from the front), you would gain additional information. While shadow images are sufficient in many applications, far more sophisticated images can be produced with modern technology. [link] shows the use of a computed tomography (CT) scanner, also called computed axial tomography (CAT) scanner. X rays are passed through a narrow section (called a slice) of the patient's body (or body part) over a range of directions. An array of many detectors on the other side of the patient registers the x rays. The system is then rotated around the patient and another image is taken, and so on. The x-ray tube and detector array are mechanically attached and so rotate together. Complex computer image processing of the relative absorption of the x rays along different directions produces a highly-detailed image. Different slices are taken as the patient moves through the scanner on a table. Multiple images of different slices can also be computer analyzed to produce three-dimensional information, sometimes enhancing specific types of tissue, as shown in [link]. G. Hounsfield (UK) and A. Cormack (US) won the Nobel Prize in Medicine in 1979 for their development of computed tomography.

A patient being positioned in a CT scanner aboard the hospital ship USNS Mercy. The CT scanner passes x rays through slices of the patient's body (or body part) over a range of directions. The relative absorption of the x rays along different directions is computer analyzed to produce highly detailed images. Three-dimensional information can be obtained from multiple slices. (credit: Rebecca Moat, U.S. Navy)

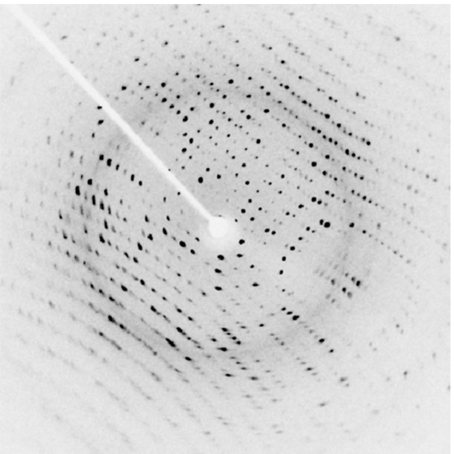


This three-dimensional image of a skull was produced by computed tomography, involving analysis of several x-ray slices of the head. (credit: Emailshankar, Wikimedia Commons)

X-Ray Diffraction and Crystallography

Since x-ray photons are very energetic, they have relatively short wavelengths. For example, the 54.4-keV K_{α} x ray of [link] has a wavelength $\lambda = \text{hc}/E = 0.0228 \text{ nm}$. Thus, typical x-ray photons act like rays when they encounter macroscopic objects, like teeth, and produce sharp shadows; however, since atoms are on the order of 0.1 nm in size, x rays can be used to detect the location, shape, and size of atoms and molecules. The process is called **x-ray diffraction**, because it involves the diffraction and interference of x rays to produce patterns that can be analyzed for information about the structures that scattered the x rays. Perhaps the most famous example of x-ray diffraction is the discovery of the double-helix structure of DNA in 1953 by an international team of scientists working at the Cavendish Laboratory—American James Watson, Englishman Francis Crick, and New Zealand–born Maurice Wilkins. Using x-ray diffraction data produced by Rosalind Franklin, they were the first to discern the structure of DNA that is so crucial to life. For this, Watson, Crick, and Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine. There is much debate and controversy over the issue that Rosalind Franklin was not included in the prize.

[link] shows a diffraction pattern produced by the scattering of x rays from a crystal. This process is known as x-ray crystallography because of the information it can yield about crystal structure, and it was the type of data Rosalind Franklin supplied to Watson and Crick for DNA. Not only do x rays confirm the size and shape of atoms, they give information on the atomic arrangements in materials. For example, current research in high-temperature superconductors involves complex materials whose lattice arrangements are crucial to obtaining a superconducting material. These can be studied using x-ray crystallography.



X-ray diffraction from the crystal of a protein, hen egg lysozyme, produced this interference pattern. Analysis of the pattern yields information about the structure of the protein. (credit: Del45, Wikimedia Commons)

Historically, the scattering of x rays from crystals was used to prove that x rays are energetic EM waves. This was suspected from the time of the discovery of x rays in 1895, but it was not until 1912 that the German Max von Laue (1879–1960) convinced two of his colleagues to scatter x rays from crystals. If a diffraction pattern is obtained, he reasoned, then the x rays must be waves, and their wavelength could be determined. (The spacing of atoms in various crystals was reasonably well known at the time, based on good values for Avogadro’s number.) The experiments were convincing, and the 1914 Nobel Prize in Physics was given to von Laue for his suggestion leading to the proof that x rays are EM waves. In 1915, the unique father-and-son team of Sir William Henry Bragg and his son Sir William Lawrence Bragg were awarded a joint Nobel Prize for inventing the x-ray spectrometer and the then-new science of x-ray analysis. The elder Bragg had migrated to Australia from England just after graduating in mathematics. He learned physics and chemistry during his career at the University of Adelaide. The younger Bragg was born in Adelaide but went back to the Cavendish Laboratories in England to a career in x-ray and neutron crystallography; he provided support for Watson, Crick, and Wilkins for their work on unraveling the mysteries of DNA and to Max Perutz for his 1962 Nobel Prize-winning work on the structure of hemoglobin. Here again, we witness the enabling nature of physics—establishing instruments and designing experiments as well as solving mysteries in the biomedical sciences.

Certain other uses for x rays will be studied in later chapters. X rays are useful in the treatment of cancer because of the inhibiting effect they have on cell reproduction. X rays observed coming from outer space are useful in determining the nature of their sources, such as neutron stars and possibly black holes. Created in nuclear bomb explosions, x rays can also be used to detect clandestine atmospheric tests of these weapons. X rays can cause excitations of atoms, which then fluoresce (emitting characteristic EM radiation), making x-ray-induced fluorescence a valuable analytical tool in a range of fields from art to archaeology.

Section Summary

- X rays are relatively high-frequency EM radiation. They are produced by transitions between inner-shell electron levels, which produce x rays characteristic of the atomic element, or by decelerating electrons.
- X rays have many uses, including medical diagnostics and x-ray diffraction.

Conceptual Questions

Exercise:

Problem:

Explain why characteristic x rays are the most energetic in the EM emission spectrum of a given element.

Exercise:

Problem:

Why does the energy of characteristic x rays become increasingly greater for heavier atoms?

Exercise:

Problem:

Observers at a safe distance from an atmospheric test of a nuclear bomb feel its heat but receive none of its copious x rays. Why is air opaque to x rays but transparent to infrared?

Exercise:

Problem:

Lasers are used to burn and read CDs. Explain why a laser that emits blue light would be capable of burning and reading more information than one that emits infrared.

Exercise:

Problem:

Crystal lattices can be examined with x rays but not UV. Why?

Exercise:

Problem:

CT scanners do not detect details smaller than about 0.5 mm. Is this limitation due to the wavelength of x rays? Explain.

Problem Exercises

Exercise:

Problem:

(a) What is the shortest-wavelength x-ray radiation that can be generated in an x-ray tube with an applied voltage of 50.0 kV? (b) Calculate the photon energy in eV. (c) Explain the relationship of the photon energy to the applied voltage.

Solution:

(a) $0.248 \times 10^{-10}\ \text{m}$

(b) 50.0 keV

(c) The photon energy is simply the applied voltage times the electron charge, so the value of the voltage in volts is the same as the value of the energy in electron volts.

Exercise:

Problem:

A color television tube also generates some x rays when its electron beam strikes the screen. What is the shortest wavelength of these x rays, if a 30.0-kV potential is used to accelerate the electrons? (Note that TVs have shielding to prevent these x rays from exposing viewers.)

Exercise:

Problem:

An x ray tube has an applied voltage of 100 kV. (a) What is the most energetic x-ray photon it can produce? Express your answer in electron volts and joules. (b) Find the wavelength of such an X–ray.

Solution:

(a) $100 \times 10^{3}\ \text{eV}$, $1.60 \times 10^{-14}\ \text{J}$

(b) 0.124×10^{-10} m

Exercise:

Problem:

The maximum characteristic x-ray photon energy comes from the capture of a free electron into a K shell vacancy. What is this photon energy in keV for tungsten, assuming the free electron has no initial kinetic energy?

Exercise:

Problem:

What are the approximate energies of the K_α and K_β x rays for copper?

Solution:

(a) 8.00 keV

(b) 9.48 keV

Glossary

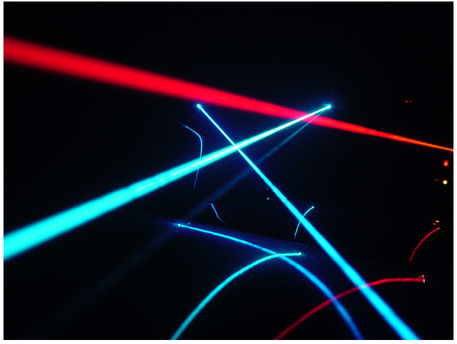
x rays
: a form of electromagnetic radiation

x-ray diffraction
: a technique that provides the detailed information about crystallographic structure of natural and manufactured materials

Applications of Atomic Excitations and De-Excitations

- Define and discuss fluorescence.
- Define metastable.
- Describe how laser emission is produced.
- Explain population inversion.
- Define and discuss holography.

Many properties of matter and phenomena in nature are directly related to atomic energy levels and their associated excitations and de-excitations. The color of a rose, the output of a laser, and the transparency of air are but a few examples. (See [link].) While it may not appear that glow-in-the-dark pajamas and lasers have much in common, they are in fact different applications of similar atomic de-excitations.

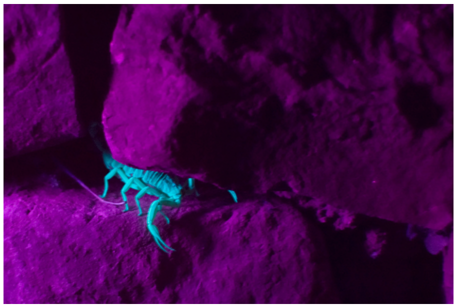


Light from a laser is based on a particular type of atomic de-excitation. (credit: Jeff Keyzer)

The color of a material is due to the ability of its atoms to absorb certain wavelengths while reflecting or reemitting others. A simple red material, for example a tomato, absorbs all visible wavelengths except red. This is because the atoms of its hydrocarbon pigment (lycopene) have levels separated by a variety of energies corresponding to all visible photon energies except red. Air is another interesting example. It is transparent to visible light, because there are few energy levels that visible photons can excite in air molecules and atoms. Visible light, thus, cannot be absorbed. Furthermore, visible light is only weakly scattered by air, because visible wavelengths are so much greater than the sizes of the air molecules and atoms. Light must pass through kilometers of air to scatter enough to cause red sunsets and blue skies.

Fluorescence and Phosphorescence

The ability of a material to emit various wavelengths of light is similarly related to its atomic energy levels. [link] shows a scorpion illuminated by a UV lamp, sometimes called a black light. Some rocks also glow in black light, the particular colors being a function of the rock’s mineral composition. Black lights are also used to make certain posters glow.



Objects glow in the visible spectrum when illuminated by an ultraviolet (black) light. Emissions are characteristic of the mineral involved, since they are related to its energy levels. In the case of scorpions, proteins near the surface of their skin give off the characteristic blue glow. This is a colorful example of fluorescence in which excitation is induced by UV radiation while de-excitation occurs in the form of visible light. (credit: Ken Bosma, Flickr)

In the fluorescence process, an atom is excited to a level several steps above its ground state by the absorption of a relatively high-energy UV photon. This is called **atomic excitation**. Once it is excited, the atom can de-excite in several ways, one of which is to re-emit a photon of the same energy as excited it, a single step back to the ground state. This is called **atomic de-excitation**. All other paths of de-excitation involve smaller steps, in which lower-energy (longer wavelength) photons are emitted. Some of these may be in the visible range, such as for the scorpion in [link]. **Fluorescence** is defined to be any process in which an atom or molecule, excited by a photon of a given energy, and de-excites by emission of a lower-energy photon.

Fluorescence can be induced by many types of energy input. Fluorescent paint, dyes, and even soap residues in clothes make colors seem brighter in sunlight by converting some UV into visible light. X rays can induce fluorescence, as is done in x-ray fluoroscopy to make brighter visible images. Electric discharges can induce fluorescence, as in so-called neon lights and in gas-discharge tubes that produce atomic and molecular spectra. Common fluorescent lights use an electric discharge in mercury vapor to cause atomic emissions from mercury atoms. The inside of a fluorescent light is coated with a fluorescent material that emits visible light over a broad spectrum of wavelengths. By choosing an appropriate coating, fluorescent lights can be made more like sunlight or like the reddish glow of candlelight, depending on needs. Fluorescent lights are more efficient in converting electrical energy into visible light than incandescent filaments (about four times as efficient), the blackbody radiation of which is primarily in the infrared due to temperature limitations.

This atom is excited to one of its higher levels by absorbing a UV photon. It can de-excite in a single step, re-emitting a photon of the same energy, or in several steps. The process is called fluorescence if the atom de-excites in smaller steps, emitting energy different from that which excited it. Fluorescence can be induced by a variety of energy inputs, such as UV, x-rays, and electrical discharge.

The spectacular Waitomo caves on North Island in New Zealand provide a natural habitat for glow-worms. The glow-worms hang up to 70 silk threads of about 30 or 40 cm each to trap prey that fly towards them in the dark. The fluorescence process is very efficient, with nearly 100% of the energy input turning into light. (In comparison, fluorescent lights are about 20% efficient.)

Fluorescence has many uses in biology and medicine. It is commonly used to label and follow a molecule within a cell. Such tagging allows one to study the structure of DNA and proteins. Fluorescent dyes and antibodies are usually used to tag the molecules, which are then illuminated with UV light and their emission of visible light is observed. Since the fluorescence of each element is characteristic, identification of elements within a sample can be done this way.

[link] shows a commonly used fluorescent dye called fluorescein. Below that, [link] reveals the diffusion of a fluorescent dye in water by observing it under UV light.

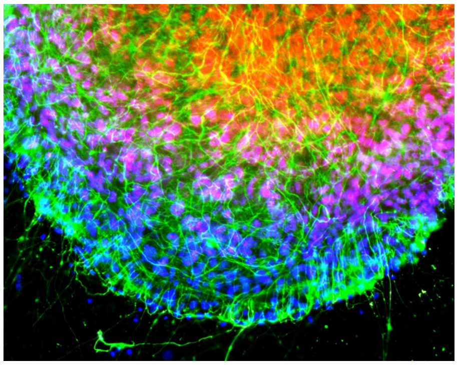
Fluorescein, shown here in powder form, is used to dye laboratory samples. (credit: Benjah-bmm27, Wikimedia Commons)

Here, fluorescent powder is added to a beaker of water. The mixture gives off a bright glow under ultraviolet light. (credit: Bricksnite, Wikimedia Commons)

Note:

Nano-Crystals

Recently, a new class of fluorescent materials has appeared—“nano-crystals.” These are single-crystal molecules less than 100 nm in size. The smallest of these are called “quantum dots.” These semiconductor indicators are very small (2–6 nm) and provide improved brightness. They also have the advantage that all colors can be excited with the same incident wavelength. They are brighter and more stable than organic dyes and have a longer lifetime than conventional phosphors. They have become an excellent tool for long-term studies of cells, including migration and morphology. ([link].)



Microscopic image of chicken cells using nano-crystals of a fluorescent dye. Cell nuclei exhibit blue fluorescence while neurofilaments exhibit green. (credit: Weerapong Prasongchean, Wikimedia Commons)

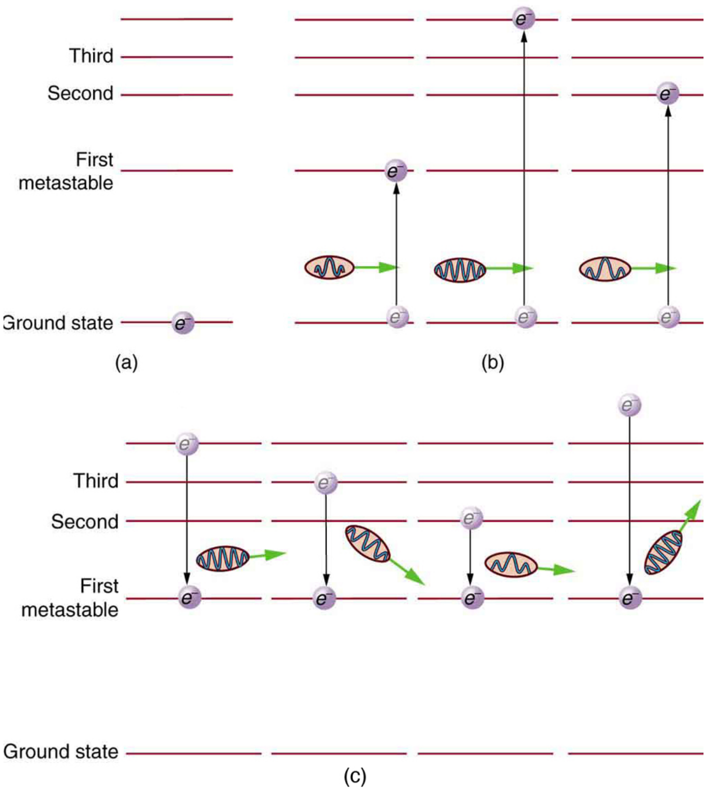
Once excited, an atom or molecule will usually spontaneously de-excite quickly. (The electrons raised to higher levels are attracted to lower ones by the positive charge of the nucleus.) Spontaneous de-excitation has a very short mean lifetime of typically about 10^{-8} s. However, some levels have significantly longer lifetimes, ranging up to milliseconds to minutes or even hours. These energy levels are inhibited and are slow in de-exciting because their quantum numbers differ greatly from those of available lower levels. Although these level lifetimes are short in human terms, they are many orders of magnitude longer than is typical and, thus, are said to be **metastable**, meaning relatively stable. **Phosphorescence** is the de-excitation of a metastable state. Glow-in-the-dark materials, such as luminous dials on some watches and clocks and on children’s toys and pajamas, are made of phosphorescent substances. Visible light excites the atoms or molecules to metastable states that decay slowly, releasing the stored excitation energy partially as visible light. In some ceramics, atomic excitation energy can be frozen in after the ceramic has cooled from its firing. It is very slowly released, but the ceramic can be induced to phosphoresce by heating—a process called “thermoluminescence.” Since the release is slow, thermoluminescence can be used to date antiquities. The less light emitted, the older the ceramic. (See [link].)

Atoms frozen in an excited state when this Chinese ceramic figure was fired can be stimulated to de-excite and emit EM radiation by heating a sample of the ceramic—a process called thermoluminescence. Since the states slowly de-excite over centuries, the amount of thermoluminescence decreases with age, making it possible to use this effect to date and authenticate antiquities. This figure dates from the 11th century. (credit: Vassil, Wikimedia Commons)

Lasers

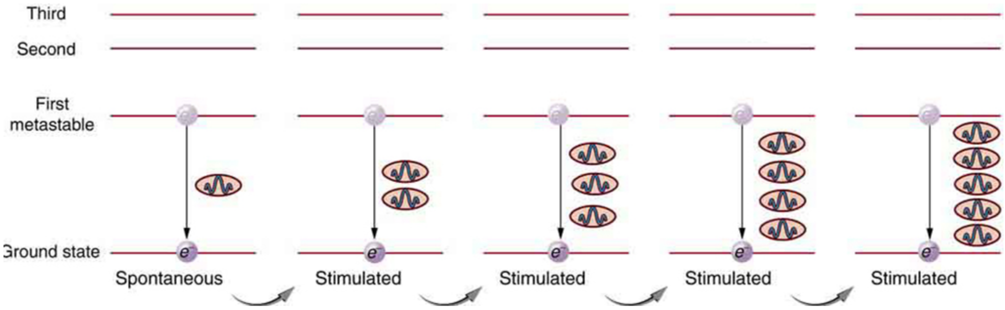
Lasers today are commonplace. Lasers are used to read bar codes at stores and in libraries, laser shows are staged for entertainment, laser printers produce high-quality images at relatively low cost, and lasers send prodigious numbers of telephone messages through optical fibers. Among other things, lasers are also employed in surveying, weapons guidance, tumor eradication, retinal welding, and for reading music CDs and computer CD-ROMs.

Why do lasers have so many varied applications? The answer is that lasers produce single-wavelength EM radiation that is also very coherent—that is, the emitted photons are in phase. Laser output can, thus, be more precisely manipulated than incoherent mixed-wavelength EM radiation from other sources. The reason laser output is so pure and coherent is based on how it is produced, which in turn depends on a metastable state in the lasing material. Suppose a material had the energy levels shown in [link]. When energy is put into a large collection of these atoms, electrons are raised to all possible levels. Most return to the ground state in less than about 10^{-8} s, but those in the metastable state linger. This includes those electrons originally excited to the metastable state and those that fell into it from above. It is possible to get a majority of the atoms into the metastable state, a condition called a **population inversion**.



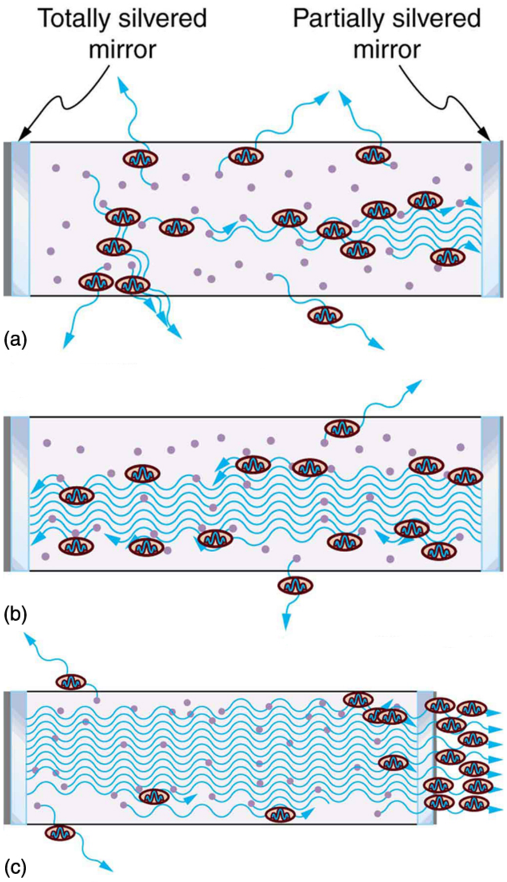
(a) Energy-level diagram for an atom showing the first few states, one of which is metastable. (b) Massive energy input excites atoms to a variety of states. (c) Most states decay quickly, leaving electrons only in the metastable and ground state. If a majority of electrons are in the metastable state, a population inversion has been achieved.

Once a population inversion is achieved, a very interesting thing can happen, as shown in [link]. An electron spontaneously falls from the metastable state, emitting a photon. This photon finds another atom in the metastable state and stimulates it to decay, emitting a second photon of *the same wavelength and in phase* with the first, and so on. **Stimulated emission** is the emission of electromagnetic radiation in the form of photons of a given frequency, triggered by photons of the same frequency. For example, an excited atom, with an electron in an energy orbit higher than normal, releases a photon of a specific frequency when the electron drops back to a lower energy orbit. If this photon then strikes another electron in the same high-energy orbit in another atom, another photon of the same frequency is released. The emitted photons and the triggering photons are always in phase, have the same polarization, and travel in the same direction. The probability of absorption of a photon is the same as the probability of stimulated emission, and so a majority of atoms must be in the metastable state to produce energy. Einstein (again Einstein, and back in 1917!) was one of the important contributors to the understanding of stimulated emission of radiation. Among other things, Einstein was the first to realize that stimulated emission and absorption are equally probable. The laser acts as a temporary energy storage device that subsequently produces a massive energy output of single-wavelength, in-phase photons.



One atom in the metastable state spontaneously decays to a lower level, producing a photon that goes on to stimulate another atom to de-excite. The second photon has exactly the same energy and wavelength as the first and is in phase with it. Both go on to stimulate the emission of other photons. A population inversion is necessary for there to be a net production rather than a net absorption of the photons.

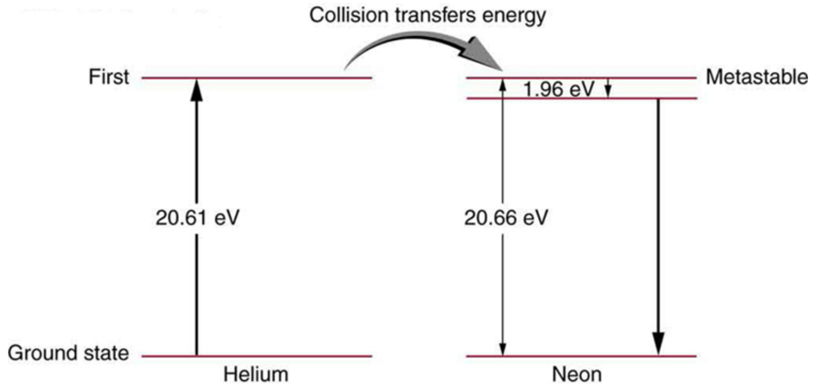
The name **laser** is an acronym for light amplification by stimulated emission of radiation, the process just described. The process was proposed and developed following the advances in quantum physics. A joint Nobel Prize was awarded in 1964 to American Charles Townes (1915–), and Nikolay Basov (1922–2001) and Aleksandr Prokhorov (1916–2002), from the Soviet Union, for the development of lasers. The Nobel Prize in 1981 went to Arthur Schawlow (1921-1999) for pioneering laser applications. The original devices were called masers, because they produced microwaves. The first working laser was created in 1960 at Hughes Research labs (CA) by T. Maiman. It used a pulsed high-powered flash lamp and a ruby rod to produce red light. Today the name laser is used for all such devices developed to produce a variety of wavelengths, including microwave, infrared, visible, and ultraviolet radiation. [link] shows how a laser can be constructed to enhance the stimulated emission of radiation. Energy input can be from a flash tube, electrical discharge, or other sources, in a process sometimes called optical pumping. A large percentage of the original pumping energy is dissipated in other forms, but a population inversion must be achieved. Mirrors can be used to enhance stimulated emission by multiple passes of the radiation back and forth through the lasing material. One of the mirrors is semitransparent to allow some of the light to pass through. The laser output from a laser is a mere 1% of the light passing back and forth in a laser.



Typical laser construction has a method of pumping energy into the lasing material to produce a population inversion. (a) Spontaneous emission begins with some photons escaping and others stimulating further emissions. (b) and (c)

Mirrors are used to enhance the probability of stimulated emission by passing photons through the material several times.

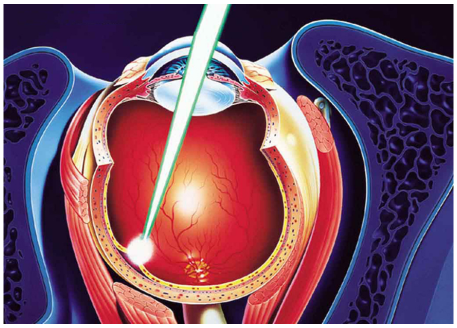
Lasers are constructed from many types of lasing materials, including gases, liquids, solids, and semiconductors. But all lasers are based on the existence of a metastable state or a phosphorescent material. Some lasers produce continuous output; others are pulsed in bursts as brief as 10^{-14} s. Some laser outputs are fantastically powerful—some greater than 10^{12} W—but the more common, everyday lasers produce something on the order of 10^{-3} W. The helium-neon laser that produces a familiar red light is very common. [link] shows the energy levels of helium and neon, a pair of noble gases that work well together. An electrical discharge is passed through a helium-neon gas mixture in which the number of atoms of helium is ten times that of neon. The first excited state of helium is metastable and, thus, stores energy. This energy is easily transferred by collision to neon atoms, because they have an excited state at nearly the same energy as that in helium. That state in neon is also metastable, and this is the one that produces the laser output. (The most likely transition is to the nearby state, producing 1.96 eV photons, which have a wavelength of 633 nm and appear red.) A population inversion can be produced in neon, because there are so many more helium atoms and these put energy into the neon. Helium-neon lasers often have continuous output, because the population inversion can be maintained even while lasing occurs. Probably the most common lasers in use today, including the common laser pointer, are semiconductor or diode lasers, made of silicon. Here, energy is pumped into the material by passing a current in the device to excite the electrons. Special coatings on the ends and fine cleavings of the semiconductor material allow light to bounce back and forth and a tiny fraction to emerge as laser light. Diode lasers can usually run continually and produce outputs in the milliwatt range.



Energy levels in helium and neon. In the common helium-neon laser, an electrical discharge pumps energy into the metastable states of both atoms. The gas mixture has about ten times more helium atoms than neon atoms. Excited helium atoms easily de-excite by transferring energy to neon in a collision. A population inversion in neon is achieved, allowing lasing by the neon to occur.

There are many medical applications of lasers. Lasers have the advantage that they can be focused to a small spot. They also have a well-defined wavelength. Many types of lasers are available today that provide wavelengths from the ultraviolet to the infrared. This is important, as one needs to be able to select a wavelength that will be preferentially absorbed by the material of interest. Objects appear a certain color because they absorb all other visible colors incident upon them. What wavelengths are absorbed depends upon the energy spacing between electron orbitals in that molecule. Unlike the hydrogen atom, biological molecules are complex and have a variety of absorption wavelengths or lines. But these can be determined and used in the selection of a laser with the appropriate wavelength. Water is transparent to the visible spectrum but will absorb light in the UV and IR regions. Blood (hemoglobin) strongly reflects red but absorbs most strongly in the UV.

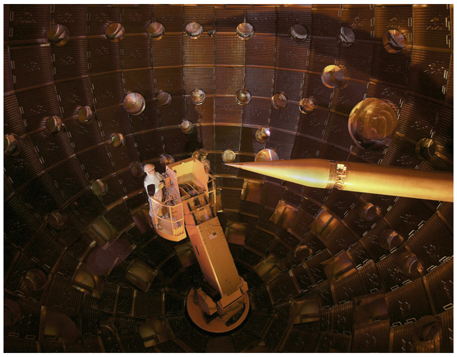
Laser surgery uses a wavelength that is strongly absorbed by the tissue it is focused upon. One example of a medical application of lasers is shown in [link]. A detached retina can result in total loss of vision. Burns made by a laser focused to a small spot on the retina form scar tissue that can hold the retina in place, salvaging the patient’s vision. Other light sources cannot be focused as precisely as a laser due to refractive dispersion of different wavelengths. Similarly, laser surgery in the form of cutting or burning away tissue is made more accurate because laser output can be very precisely focused and is preferentially absorbed because of its single wavelength. Depending upon what part or layer of the retina needs repairing, the appropriate type of laser can be selected. For the repair of tears in the retina, a green argon laser is generally used. This light is absorbed well by tissues containing blood, so coagulation or “welding” of the tear can be done.



A detached retina is burned by a laser designed to focus on a small spot on the retina, the resulting scar tissue holding it in place. The lens of the eye is used to focus the light, as is the device bringing the laser output to the eye.

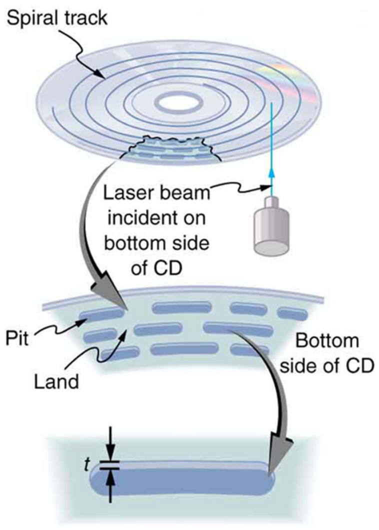
In dentistry, the use of lasers is rising. Lasers are most commonly used for surgery on the soft tissue of the mouth. They can be used to remove ulcers, stop bleeding, and reshape gum tissue. Their use in cutting into bones and teeth is not quite so common; here the erbium YAG (yttrium aluminum garnet) laser is used.

The massive combination of lasers shown in [link] can be used to induce nuclear fusion, the energy source of the sun and hydrogen bombs. Since lasers can produce very high power in very brief pulses, they can be used to focus an enormous amount of energy on a small glass sphere containing fusion fuel. Not only does the incident energy increase the fuel temperature significantly so that fusion can occur, it also compresses the fuel to great density, enhancing the probability of fusion. The compression or implosion is caused by the momentum of the impinging laser photons.



This system of lasers at Lawrence Livermore Laboratory is used to ignite nuclear fusion. A tremendous burst of energy is focused on a small fuel pellet, which is imploded to the high density and temperature needed to make the fusion reaction proceed. (credit: Lawrence Livermore National Laboratory, Lawrence Livermore National Security, LLC, and the Department of Energy)

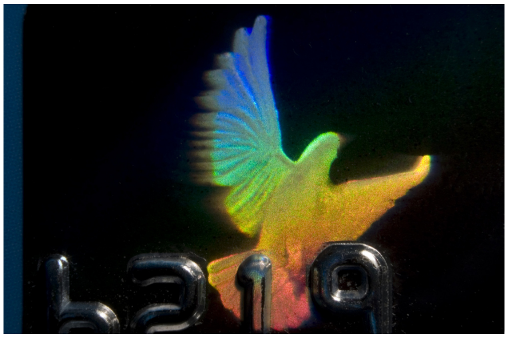
Music CDs are now so common that vinyl records are quaint antiquities. CDs (and DVDs) store information digitally and have a much larger information-storage capacity than vinyl records. An entire encyclopedia can be stored on a single CD. [link] illustrates how the information is stored and read from the CD. Pits made in the CD by a laser can be tiny and very accurately spaced to record digital information. These are read by having an inexpensive solid-state infrared laser beam scatter from pits as the CD spins, revealing their digital pattern and the information encoded upon them.



A CD has digital information

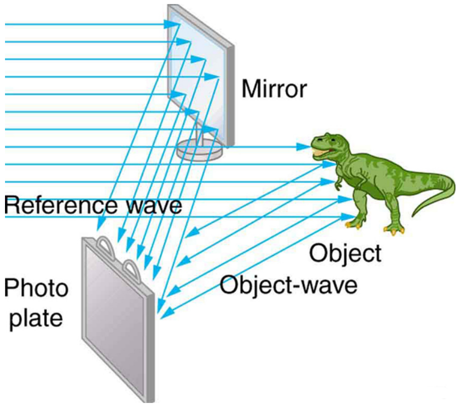
stored in the form of laser-created pits on its surface. These in turn can be read by detecting the laser light scattered from the pit. Large information capacity is possible because of the precision of the laser. Shorter-wavelength lasers enable greater storage capacity.

Holograms, such as those in [link], are true three-dimensional images recorded on film by lasers. Holograms are used for amusement, decoration on novelty items and magazine covers, security on credit cards and driver’s licenses (a laser and other equipment is needed to reproduce them), and for serious three-dimensional information storage. You can see that a hologram is a true three-dimensional image, because objects change relative position in the image when viewed from different angles.



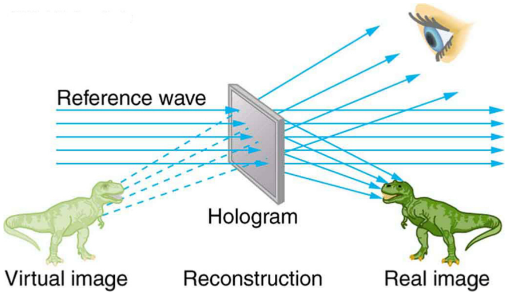
Credit cards commonly have holograms for logos, making them difficult to reproduce (credit: Dominic Alves, Flickr)

The name **hologram** means “entire picture” (from the Greek *holo*, as in holistic), because the image is three-dimensional. **Holography** is the process of producing holograms and, although they are recorded on photographic film, the process is quite different from normal photography. Holography uses light interference or wave optics, whereas normal photography uses geometric optics. [link] shows one method of producing a hologram. Coherent light from a laser is split by a mirror, with part of the light illuminating the object. The remainder, called the reference beam, shines directly on a piece of film. Light scattered from the object interferes with the reference beam, producing constructive and destructive interference. As a result, the exposed film looks foggy, but close examination reveals a complicated interference pattern stored on it. Where the interference was constructive, the film (a negative actually) is darkened. Holography is sometimes called lensless photography, because it uses the wave characteristics of light as contrasted to normal photography, which uses geometric optics and so requires lenses.



Production of a hologram. Single-wavelength coherent light from a laser produces a well-defined interference pattern on a piece of film. The laser beam is split by a partially silvered mirror, with part of the light illuminating the object and the remainder shining directly on the film.

Light falling on a hologram can form a three-dimensional image. The process is complicated in detail, but the basics can be understood as shown in [link], in which a laser of the same type that exposed the film is now used to illuminate it. The myriad tiny exposed regions of the film are dark and block the light, while less exposed regions allow light to pass. The film thus acts much like a collection of diffraction gratings with various spacings. Light passing through the hologram is diffracted in various directions, producing both real and virtual images of the object used to expose the film. The interference pattern is the same as that produced by the object. Moving your eye to various places in the interference pattern gives you different perspectives, just as looking directly at the object would. The image thus looks like the object and is three-dimensional like the object.



A transmission hologram is one that produces real and virtual images when a laser of the same type as that which exposed the hologram is passed through it. Diffraction from various parts of the film produces the same interference pattern as the object that was used to expose it.

The hologram illustrated in [link] is a transmission hologram. Holograms that are viewed with reflected light, such as the white light holograms on credit cards, are reflection holograms and are more common. White light holograms often appear a little blurry with rainbow edges, because the diffraction patterns of various colors of light are at slightly different locations due to their different wavelengths. Further uses of holography include all types of 3-D information storage, such as of statues in museums and engineering studies of structures and 3-D images of human organs. Invented in the late 1940s by Dennis Gabor (1900–1970), who won the

1971 Nobel Prize in Physics for his work, holography became far more practical with the development of the laser. Since lasers produce coherent single-wavelength light, their interference patterns are more pronounced. The precision is so great that it is even possible to record numerous holograms on a single piece of film by just changing the angle of the film for each successive image. This is how the holograms that move as you walk by them are produced—a kind of lensless movie.

In a similar way, in the medical field, holograms have allowed complete 3-D holographic displays of objects from a stack of images. Storing these images for future use is relatively easy. With the use of an endoscope, high-resolution 3-D holographic images of internal organs and tissues can be made.

Section Summary

- An important atomic process is fluorescence, defined to be any process in which an atom or molecule is excited by absorbing a photon of a given energy and de-excited by emitting a photon of a lower energy.
- Some states live much longer than others and are termed metastable.
- Phosphorescence is the de-excitation of a metastable state.
- Lasers produce coherent single-wavelength EM radiation by stimulated emission, in which a metastable state is stimulated to decay.
- Lasing requires a population inversion, in which a majority of the atoms or molecules are in their metastable state.

Conceptual Questions

Exercise:

Problem:

How do the allowed orbits for electrons in atoms differ from the allowed orbits for planets around the sun? Explain how the correspondence principle applies here.

Exercise:

Problem:

Atomic and molecular spectra are discrete. What does discrete mean, and how are discrete spectra related to the quantization of energy and electron orbits in atoms and molecules?

Exercise:

Problem:

Hydrogen gas can only absorb EM radiation that has an energy corresponding to a transition in the atom, just as it can only emit these discrete energies. When a spectrum is taken of the solar corona, in which a broad range of EM wavelengths are passed through very hot hydrogen gas, the absorption spectrum shows all the features of the emission spectrum. But when such EM radiation passes through room-temperature hydrogen gas, only the Lyman series is absorbed. Explain the difference.

Exercise:

Problem:

Lasers are used to burn and read CDs. Explain why a laser that emits blue light would be capable of burning and reading more information than one that emits infrared.

Exercise:

Problem:

The coating on the inside of fluorescent light tubes absorbs ultraviolet light and subsequently emits visible light. An inventor claims that he is able to do the reverse process. Is the inventor’s claim possible?

Exercise:

Problem:

What is the difference between fluorescence and phosphorescence?

Exercise:

Problem:

How can you tell that a hologram is a true three-dimensional image and that those in 3-D movies are not?

Problem Exercises

Exercise:

Problem:

[link] shows the energy-level diagram for neon. (a) Verify that the energy of the photon emitted when neon goes from its metastable state to the one immediately below is equal to 1.96 eV. (b) Show that the wavelength of this radiation is 633 nm. (c) What wavelength is emitted when the neon makes a direct transition to its ground state?

Solution:

(a) 1.96 eV

(b) $(1240\text{ eV}\cdot\text{nm})/(1.96\text{ eV})=633\text{ nm}$

(c) 60.0 nm

Exercise:

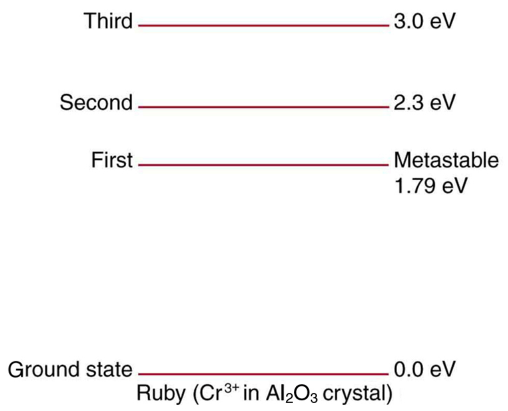
Problem:

A helium-neon laser is pumped by electric discharge. What wavelength electromagnetic radiation would be needed to pump it? See [link] for energy-level information.

Exercise:

Problem:

Ruby lasers have chromium atoms doped in an aluminum oxide crystal. The energy level diagram for chromium in a ruby is shown in [link]. What wavelength is emitted by a ruby laser?



Chromium atoms in an aluminum oxide crystal have these energy levels, one of which is metastable. This is the basis of a ruby laser. Visible light can pump the atom into an excited state above the metastable state to achieve a population inversion.

Solution:

693 nm

Exercise:

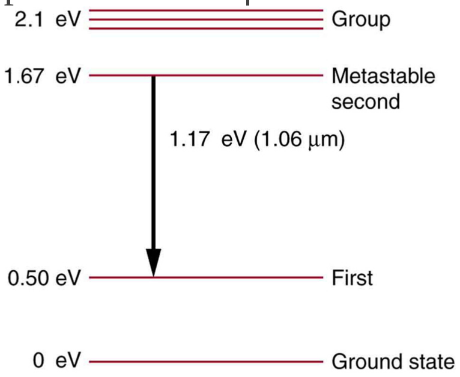
Problem:

(a) What energy photons can pump chromium atoms in a ruby laser from the ground state to its second and third excited states? (b) What are the wavelengths of these photons? Verify that they are in the visible part of the spectrum.

Exercise:

Problem:

Some of the most powerful lasers are based on the energy levels of neodymium in solids, such as glass, as shown in [link]. (a) What average wavelength light can pump the neodymium into the levels above its metastable state? (b) Verify that the 1.17 eV transition produces 1.06 μm radiation.



Neodymium atoms in glass have these energy levels, one of which is metastable. The group of levels above the metastable state is convenient for achieving a population inversion, since photons of many different energies can be absorbed by atoms in the ground state.

Solution:

(a) 590 nm

(b) $(1240\ \text{eV}\cdot\text{nm})/(1.17\ \text{eV}) = 1.06\ \mu\text{m}$

Glossary

metastable
: a state whose lifetime is an order of magnitude longer than the most short-lived states

atomic excitation
: a state in which an atom or ion acquires the necessary energy to promote one or more of its electrons to electronic states higher in energy than their ground state

atomic de-excitation
: process by which an atom transfers from an excited electronic state back to the ground state electronic configuration; often occurs by emission of a photon

laser
: acronym for light amplification by stimulated emission of radiation

phosphorescence
: the de-excitation of a metastable state

population inversion
: the condition in which the majority of atoms in a sample are in a metastable state

stimulated emission
: emission by atom or molecule in which an excited state is stimulated to decay, most readily caused by a photon of the same energy that is necessary to excite the state

hologram
: means *entire picture* (from the Greek word *holo*, as in holistic), because the image produced is three dimensional

holography
: the process of producing holograms

fluorescence

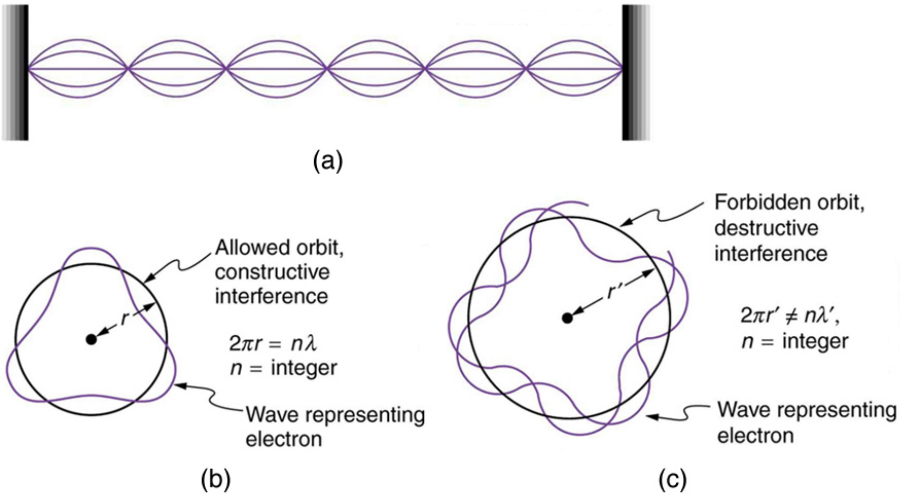
any process in which an atom or molecule, excited by a photon of a given energy, de-excites by emission of a lower-energy photon

The Wave Nature of Matter Causes Quantization

- Explain Bohr’s model of atom.
- Define and describe quantization of angular momentum.
- Calculate the angular momentum for an orbit of atom.
- Define and describe the wave-like properties of matter.

After visiting some of the applications of different aspects of atomic physics, we now return to the basic theory that was built upon Bohr’s atom. Einstein once said it was important to keep asking the questions we eventually teach children not to ask. Why is angular momentum quantized? You already know the answer. Electrons have wave-like properties, as de Broglie later proposed. They can exist only where they interfere constructively, and only certain orbits meet proper conditions, as we shall see in the next module.

Following Bohr’s initial work on the hydrogen atom, a decade was to pass before de Broglie proposed that matter has wave properties. The wave-like properties of matter were subsequently confirmed by observations of electron interference when scattered from crystals. Electrons can exist only in locations where they interfere constructively. How does this affect electrons in atomic orbits? When an electron is bound to an atom, its wavelength must fit into a small space, something like a standing wave on a string. (See [link].) Allowed orbits are those orbits in which an electron constructively interferes with itself. Not all orbits produce constructive interference. Thus only certain orbits are allowed—the orbits are quantized.


(a)
Allowed orbit, constructive interference
r
2πr = nλ
n = integer
Wave representing electron
(b)
Forbidden orbit, destructive interference
r′
2πr′ ≠ nλ′,
n = integer
Wave representing electron
(c)

(a) Waves on a string have a wavelength related to the length of the string, allowing them to interfere constructively. (b) If we imagine the string bent into a closed circle, we get a rough idea of how electrons in circular orbits can interfere constructively. (c) If the wavelength does not fit into the circumference, the electron interferes destructively; it cannot exist in such an orbit.

For a circular orbit, constructive interference occurs when the electron's wavelength fits neatly into the circumference, so that wave crests always align with crests and wave troughs align with troughs, as shown in [link] (b). More precisely, when an integral multiple of the electron's wavelength equals the circumference of the orbit, constructive interference is obtained. In equation form, the *condition for constructive interference and an allowed electron orbit* is

Equation:

$$n\lambda_n = 2\pi r_n (n = 1, 2, 3 \text{...}),$$

where λ_n is the electron's wavelength and r_n is the radius of that circular orbit. The de Broglie wavelength is $\lambda = h/p = h/\ mv$, and so here $\lambda = h/m_e v$. Substituting this into the previous condition for constructive interference produces an interesting result:

Equation:

$$\frac{\text{nh}}{m_e v} = 2\text{π}r_n.$$

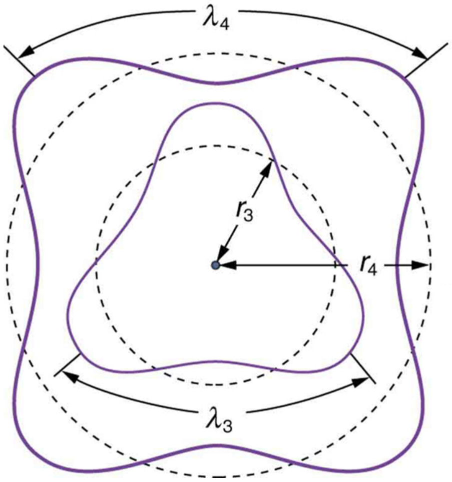
Rearranging terms, and noting that $L = \text{mvr}$ for a circular orbit, we obtain the quantization of angular momentum as the condition for allowed orbits:

Equation:

$$L = m_e \mathrm{vr}_n = n\frac{h}{2\pi}(n = 1, 2, 3 \ldots).$$

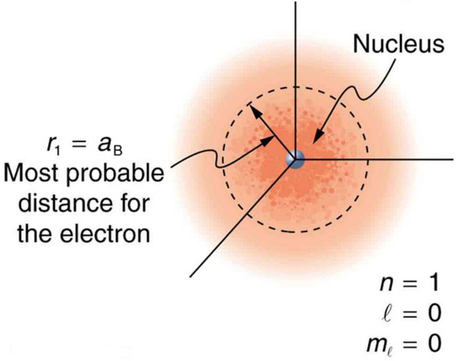
This is what Bohr was forced to hypothesize as the rule for allowed orbits, as stated earlier. We now realize that it is the condition for constructive interference of an electron in a circular orbit. [link] illustrates this for $n = 3$ and $n = 4$.

Note:
Waves and Quantization
The wave nature of matter is responsible for the quantization of energy levels in bound systems. Only those states where matter interferes constructively exist, or are “allowed.” Since there is a lowest orbit where this is possible in an atom, the electron cannot spiral into the nucleus. It cannot exist closer to or inside the nucleus. The wave nature of matter is what prevents matter from collapsing and gives atoms their sizes.



The third and fourth allowed circular orbits have three and four wavelengths, respectively, in their circumferences.

Because of the wave character of matter, the idea of well-defined orbits gives way to a model in which there is a cloud of probability, consistent with Heisenberg’s uncertainty principle. [link] shows how this applies to the ground state of hydrogen. If you try to follow the electron in some well-defined orbit using a probe that has a small enough wavelength to get some details, you will instead knock the electron out of its orbit. Each measurement of the electron’s position will find it to be in a definite location somewhere near the nucleus. Repeated measurements reveal a cloud of probability like that in the figure, with each speck the location determined by a single measurement. There is not a well-defined, circular-orbit type of distribution. Nature again proves to be different on a small scale than on a macroscopic scale.



The ground state of a hydrogen atom has a probability cloud describing the position of its electron. The probability of finding the electron is proportional to the darkness of the cloud. The electron can be closer or farther than the Bohr radius, but it is very unlikely to be a great distance from the nucleus.

There are many examples in which the wave nature of matter causes quantization in bound systems such as the atom. Whenever a particle is confined or bound to a small space, its allowed wavelengths are those which fit into that space. For example, the particle in a box model describes a particle free to move in a small space surrounded by impenetrable barriers. This is true in blackbody radiators (atoms and molecules) as well as in atomic and molecular spectra. Various atoms and molecules will have different sets of electron orbits, depending on the size and complexity of the system. When a system is large, such as a grain of sand, the tiny particle waves in it can fit in so many ways that it becomes impossible to see that the allowed states are discrete. Thus the correspondence principle is satisfied. As systems become large, they gradually look less grainy, and quantization becomes less evident. Unbound systems (small or not), such as an electron freed from an atom, do not have quantized energies, since their wavelengths are not constrained to fit in a certain volume.

Note:
PhET Explorations: Quantum Wave Interference
When do photons, electrons, and atoms behave like particles and when do they behave like waves? Watch waves spread out and interfere as they pass through a double slit, then get detected on a screen as tiny dots. Use quantum detectors to explore how measurements change the waves and the patterns they produce on the screen.

Quantum Wave

Interference

Section Summary

- Quantization of orbital energy is caused by the wave nature of matter. Allowed orbits in atoms occur for constructive interference of electrons in the orbit, requiring an integral number of wavelengths to fit in an orbit’s circumference; that is,
  Equation:
  $$n\lambda_n = 2\pi r_n (n = 1, 2, 3 \ldots),$$
  where λ_n is the electron’s de Broglie wavelength.
- Owing to the wave nature of electrons and the Heisenberg uncertainty principle, there are no well-defined orbits; rather, there are clouds of probability.
- Bohr correctly proposed that the energy and radii of the orbits of electrons in atoms are quantized, with energy for transitions between orbits given by
  Equation:
  $$\Delta E = hf = E_{\mathrm{i}} - E_{\mathrm{f}},$$
  where ΔE is the change in energy between the initial and final orbits and hf is the energy of an absorbed or emitted photon.
- It is useful to plot orbit energies on a vertical graph called an energy-level diagram.
- The allowed orbits are circular, Bohr proposed, and must have quantized orbital angular momentum given by
  Equation:
  $$L = m_e v r_n = n \frac{h}{2\pi} (n = 1, 2, 3 \ldots),$$

where L is the angular momentum, r_n is the radius of orbit n, and h is Planck’s constant.

Conceptual Questions

Exercise:

Problem:

How is the de Broglie wavelength of electrons related to the quantization of their orbits in atoms and molecules?

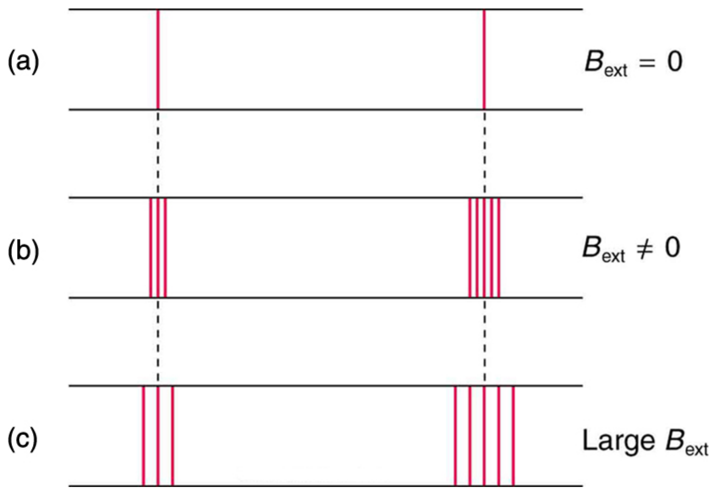
Patterns in Spectra Reveal More Quantization

- State and discuss the Zeeman effect.
- Define orbital magnetic field.
- Define orbital angular momentum.
- Define space quantization.

High-resolution measurements of atomic and molecular spectra show that the spectral lines are even more complex than they first appear. In this section, we will see that this complexity has yielded important new information about electrons and their orbits in atoms.

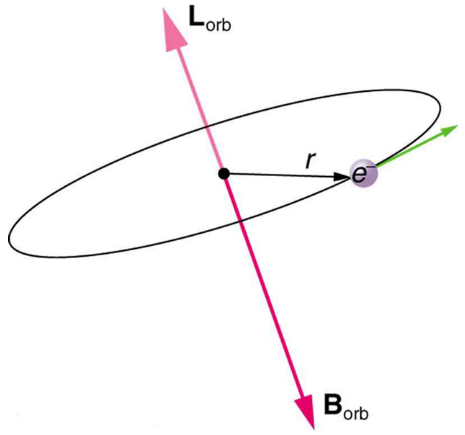
In order to explore the substructure of atoms (and knowing that magnetic fields affect moving charges), the Dutch physicist Hendrik Lorentz (1853–1930) suggested that his student Pieter Zeeman (1865–1943) study how spectra might be affected by magnetic fields. What they found became known as the **Zeeman effect**, which involved spectral lines being split into two or more separate emission lines by an external magnetic field, as shown in [link]. For their discoveries, Zeeman and Lorentz shared the 1902 Nobel Prize in Physics.

Zeeman splitting is complex. Some lines split into three lines, some into five, and so on. But one general feature is that the amount the split lines are separated is proportional to the applied field strength, indicating an interaction with a moving charge. The splitting means that the quantized energy of an orbit is affected by an external magnetic field, causing the orbit to have several discrete energies instead of one. Even without an external magnetic field, very precise measurements showed that spectral lines are doublets (split into two), apparently by magnetic fields within the atom itself.

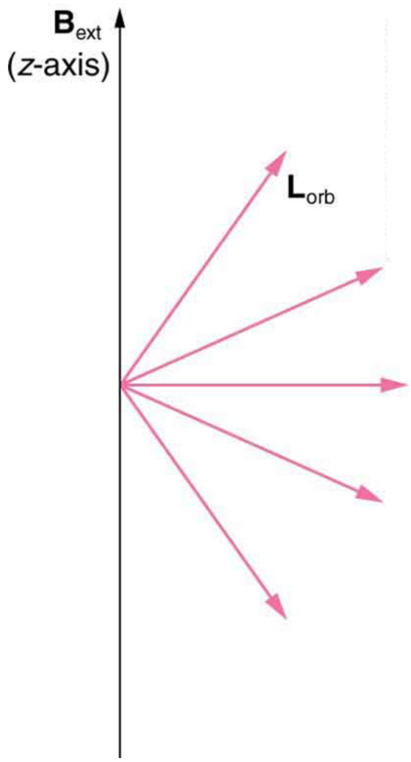


The Zeeman effect is the splitting of spectral lines when a magnetic field is applied. The number of lines formed varies, but the spread is proportional to the strength of the applied field. (a) Two spectral lines with no external magnetic field. (b) The lines split when the field is applied. (c) The splitting is greater when a stronger field is applied.

Bohr’s theory of circular orbits is useful for visualizing how an electron’s orbit is affected by a magnetic field. The circular orbit forms a current loop, which creates a magnetic field of its own, as seen in [link]. Note that the **orbital magnetic field** and the **orbital angular momentum** are along the same line. The external magnetic field and the orbital magnetic field interact; a torque is exerted to align them. A torque rotating a system through some angle does work so that there is energy associated with this interaction. Thus, orbits at different angles to the external magnetic field have different energies. What is remarkable is that the energies are quantized—the magnetic field splits the spectral lines into several discrete lines that have different energies. This means that only certain angles are allowed between the orbital angular momentum and the external field, as seen in [link].



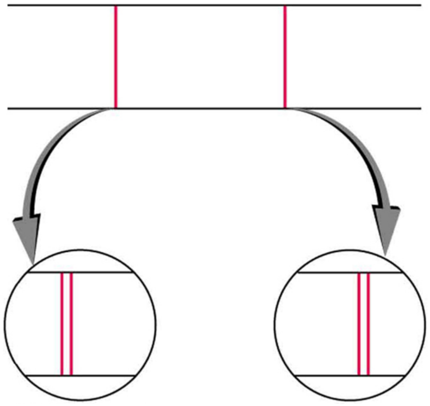
The approximate picture of an electron in a circular orbit illustrates how the current loop produces its own magnetic field, called . It also shows how is along the same line as the orbital angular momentum .



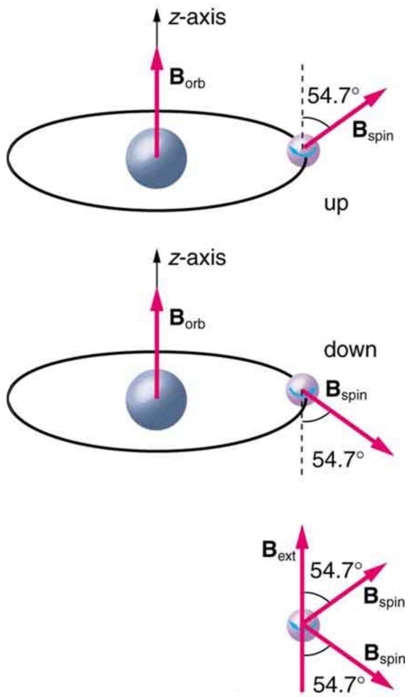
Only certain angles are allowed between the orbital angular momentum and an external magnetic field. This is implied by the fact that the Zeeman effect splits spectral lines into several discrete lines. Each line is associated with an angle between the external magnetic field and magnetic fields due to electrons and their orbits.

We already know that the magnitude of angular momentum is quantized for electron orbits in atoms. The new insight is that the *direction of the orbital angular momentum is also quantized*. The fact that the orbital angular momentum can have only certain directions is called **space quantization**. Like many aspects of quantum mechanics, this quantization of direction is totally unexpected. On the macroscopic scale, orbital angular momentum, such as that of the moon around the earth, can have any magnitude and be in any direction.

Detailed treatment of space quantization began to explain some complexities of atomic spectra, but certain patterns seemed to be caused by something else. As mentioned, spectral lines are actually closely spaced doublets, a characteristic called **fine structure**, as shown in [link]. The doublet changes when a magnetic field is applied, implying that whatever causes the doublet interacts with a magnetic field. In 1925, Sem Goudsmit and George Uhlenbeck, two Dutch physicists, successfully argued that electrons have properties analogous to a macroscopic charge spinning on its axis. Electrons, in fact, have an internal or intrinsic angular momentum called **intrinsic spin** . Since electrons are charged, their intrinsic spin creates an **intrinsic magnetic field** , which interacts with their orbital magnetic field . Furthermore, *electron intrinsic spin is quantized in magnitude and direction*, analogous to the situation for orbital angular momentum. The spin of the electron can have only one magnitude, and its direction can be at only one of two angles relative to a magnetic field, as seen in [link]. We refer to this as spin up or spin down for the electron. Each spin direction has a different energy; hence, spectroscopic lines are split into two. Spectral doublets are now understood as being due to electron spin.



Fine structure. Upon close examination, spectral lines are doublets, even in the absence of an external magnetic field. The electron has an intrinsic magnetic field that interacts with its orbital magnetic field.



z-axis
$\mathbf{B}_{\text{orb}}$
54.7°
$\mathbf{B}_{\text{spin}}$
up
z-axis
$\mathbf{B}_{\text{orb}}$
down
$\mathbf{B}_{\text{spin}}$
54.7°
$\mathbf{B}_{\text{ext}}$
54.7°
$\mathbf{B}_{\text{spin}}$
$\mathbf{B}_{\text{spin}}$
54.7°

The intrinsic magnetic field of an electron is attributed to its spin, , roughly pictured to be due to its charge spinning on its axis. This is only a crude model, since electrons seem to have no size. The spin and intrinsic magnetic field of the electron can make only one of two angles with another magnetic field, such as that created by the electron’s orbital motion. Space is quantized for spin as well as for orbital angular momentum.

These two new insights—that the direction of angular momentum, whether orbital or spin, is quantized, and that electrons have intrinsic spin—help to explain many of the complexities of atomic and molecular spectra. In magnetic resonance imaging, it is the way that the intrinsic magnetic field of hydrogen and biological atoms interact with an external field that underlies the diagnostic fundamentals.

Section Summary

- The Zeeman effect—the splitting of lines when a magnetic field is applied—is caused by other quantized entities in atoms.
- Both the magnitude and direction of orbital angular momentum are quantized.
- The same is true for the magnitude and direction of the intrinsic spin of electrons.

Conceptual Questions

Exercise:

Problem:

What is the Zeeman effect, and what type of quantization was discovered because of this effect?

Glossary

Zeeman effect

the effect of external magnetic fields on spectral lines

intrinsic spin
the internal or intrinsic angular momentum of electrons

orbital angular momentum
an angular momentum that corresponds to the quantum analog of classical angular momentum

fine structure
the splitting of spectral lines of the hydrogen spectrum when the spectral lines are examined at very high resolution

space quantization
the fact that the orbital angular momentum can have only certain directions

intrinsic magnetic field
the magnetic field generated due to the intrinsic spin of electrons

orbital magnetic field
the magnetic field generated due to the orbital motion of electrons

Quantum Numbers and Rules

- Define quantum number.
- Calculate angle of angular momentum vector with an axis.
- Define spin quantum number.

Physical characteristics that are quantized—such as energy, charge, and angular momentum—are of such importance that names and symbols are given to them. The values of quantized entities are expressed in terms of **quantum numbers**, and the rules governing them are of the utmost importance in determining what nature is and does. This section covers some of the more important quantum numbers and rules—all of which apply in chemistry, material science, and far beyond the realm of atomic physics, where they were first discovered. Once again, we see how physics makes discoveries which enable other fields to grow.

The *energy states of bound systems are quantized*, because the particle wavelength can fit into the bounds of the system in only certain ways. This was elaborated for the hydrogen atom, for which the allowed energies are expressed as $E_n \propto 1/n^2$, where $n = 1, 2, 3, \ldots$. We define n to be the principal quantum number that labels the basic states of a system. The lowest-energy state has $n = 1$, the first excited state has $n = 2$, and so on. Thus the allowed values for the principal quantum number are

Equation:

$$n = 1, 2, 3, \ldots.$$

This is more than just a numbering scheme, since the energy of the system, such as the hydrogen atom, can be expressed as some function of n, as can other characteristics (such as the orbital radii of the hydrogen atom).

The fact that the *magnitude of angular momentum is quantized* was first recognized by Bohr in relation to the hydrogen atom; it is now known to be true in general. With the development of quantum mechanics, it was found that the magnitude of angular momentum L can have only the values

Equation:

$$L = \sqrt{l(l+1)}\frac{h}{2\pi} \quad (l = 0, 1, 2, \ldots, n-1),$$

where l is defined to be the **angular momentum quantum number**. The rule for l in atoms is given in the parentheses. Given n, the value of l can be any integer from zero up to $n - 1$. For example, if $n = 4$, then l can be 0, 1, 2, or 3.

Note that for $n = 1$, l can only be zero. This means that the ground-state angular momentum for hydrogen is actually zero, not $h/2\pi$ as Bohr proposed. The picture of circular orbits is not valid, because there would be angular momentum for any circular orbit. A more valid picture is the cloud of probability shown for the ground state of hydrogen in [link]. The electron actually spends time in and near the nucleus. The reason the electron does not remain in the nucleus is related to Heisenberg's uncertainty principle—the electron's energy would have to be much too large to be confined to the small space of the nucleus. Now the first excited state of hydrogen has $n = 2$, so that l can be either 0 or 1, according to the rule in $L = \sqrt{l(l+1)}\frac{h}{2\pi}$. Similarly, for $n = 3$, l can be 0, 1, or 2. It is often most convenient to state the value of l, a simple integer, rather than calculating the value of L from $L = \sqrt{l(l+1)}\frac{h}{2\pi}$. For example, for $l = 2$, we see that

Equation:

$$L = \sqrt{2(2+1)}\frac{h}{2\pi} = \sqrt{6}\frac{h}{2\pi} = 0.390h = 2.58 \times 10^{-34}\ \text{J} \cdot \text{s}.$$

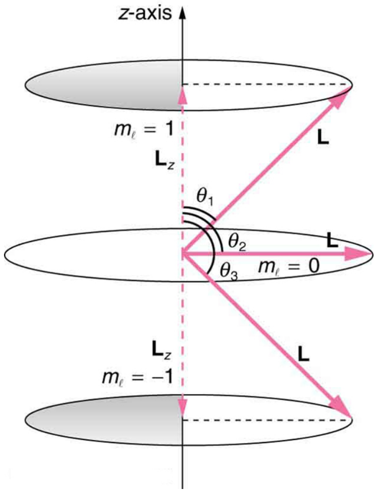
It is much simpler to state $l = 2$.

As recognized in the Zeeman effect, the *direction of angular momentum is quantized*. We now know this is true in all circumstances. It is found that the component of angular momentum along one direction in space, usually called the z-axis, can have only certain values of L_z. The direction in space must be related to something physical, such as the direction of the magnetic field at that location. This is an aspect of relativity. Direction has no meaning if there is nothing that varies with direction, as does magnetic force. The allowed values of L_z are

Equation:

$$L_z = m_l \frac{h}{2\pi} \quad (m_l = -l, -l+1, ..., \ -1, 0, 1, ... \ l-1, l),$$

where L_z is the **z-component of the angular momentum** and m_l is the angular momentum projection quantum number. The rule in parentheses for the values of m_l is that it can range from $-l$ to l in steps of one. For example, if $l = 2$, then m_l can have the five values –2, –1, 0, 1, and 2. Each m_l corresponds to a different energy in the presence of a magnetic field, so that they are related to the splitting of spectral lines into discrete parts, as discussed in the preceding section. If the z-component of angular momentum can have only certain values, then the angular momentum can have only certain directions, as illustrated in [link].



The component of a given angular momentum along the z-axis (defined by the direction of a magnetic field) can have only certain values; these are shown here for $l = 1$, for which $m_l = -1, 0,$ and $+1$.

The direction of L is quantized in the sense that it can have only certain angles relative to the z-axis.

Example:
What Are the Allowed Directions?
Calculate the angles that the angular momentum vector $\mathbf{L}$ can make with the z-axis for $l=1$, as illustrated in [link].

Strategy
[link] represents the vectors $\mathbf{L}$ and $\mathbf{L}_z$ as usual, with arrows proportional to their magnitudes and pointing in the correct directions. $\mathbf{L}$ and $\mathbf{L}_z$ form a right triangle, with $\mathbf{L}$ being the hypotenuse and $\mathbf{L}_z$ the adjacent side. This means that the ratio of $\mathbf{L}_z$ to $\mathbf{L}$ is the cosine of the angle of interest. We can find $\mathbf{L}$ and $\mathbf{L}_z$ using $L=\sqrt{l(l+1)}\frac{h}{2\pi}$ and $L_z=m\frac{h}{2\pi}$.

Solution
We are given $l=1$, so that m_l can be +1, 0, or −1. Thus L has the value given by $L=\sqrt{l(l+1)}\frac{h}{2\pi}$.

Equation:

$$L=\frac{\sqrt{l(l+1)}h}{2\pi}=\frac{\sqrt{2}h}{2\pi}$$

L_z can have three values, given by $L_z=m_l\frac{h}{2\pi}$.

Equation:

$$L_z=m_l\frac{h}{2\pi}=\begin{array}{lll}\frac{h}{2\pi}, & m_l & = & +1\\ 0, & m_l & = & 0\\ -\frac{h}{2\pi}, & m_l & = & -1\end{array}$$

As can be seen in [link], $\cos\,\theta=\mathrm{L}_z/\mathrm{L}$, and so for $m_l=+\,1$, we have

Equation:

$$\cos\,\theta_1=\frac{L_Z}{L}=\frac{\frac{h}{2\pi}}{\frac{\sqrt{2}h}{2\pi}}=\frac{1}{\sqrt{2}}=0.707.$$

Thus,

Equation:

$$\theta_1=\cos^{-1}0.707=45.0^{\mathrm{o}}.$$

Similarly, for $m_l=0$, we find $\cos\,\theta_2=0$; thus,

Equation:

$$\theta_2 = \cos^{-1}0 = 90.0^{\text{o}}.$$

And for $m_l = -1$,
Equation:

$$\cos\theta_3 = \frac{L_Z}{L} = \frac{-\frac{h}{2\pi}}{\frac{\sqrt{2}h}{2\pi}} = -\frac{1}{\sqrt{2}} = -0.707,$$

so that
Equation:

$$\theta_3 = \cos^{-1}(-0.707) = 135.0^{\text{o}}.$$

Discussion
The angles are consistent with the figure. Only the angle relative to the z-axis is quantized. L can point in any direction as long as it makes the proper angle with the z-axis. Thus the angular momentum vectors lie on cones as illustrated. This behavior is not observed on the large scale. To see how the correspondence principle holds here, consider that the smallest angle (θ_1 in the example) is for the maximum value of $m_l = 0$, namely $m_l = l$. For that smallest angle,
Equation:

$$\cos\theta = \frac{L_z}{L} = \frac{l}{\sqrt{l(l+1)}},$$

which approaches 1 as l becomes very large. If $\cos\theta = 1$, then $\theta = 0^{\text{o}}$. Furthermore, for large l, there are many values of m_l, so that all angles become possible as l gets very large.

Intrinsic Spin Angular Momentum Is Quantized in Magnitude and Direction

There are two more quantum numbers of immediate concern. Both were first discovered for electrons in conjunction with fine structure in atomic spectra. It is now well established that electrons and other fundamental particles have *intrinsic spin*, roughly analogous to a planet spinning on its axis. This spin is a fundamental characteristic of particles, and only one magnitude of intrinsic spin is allowed for a given type of particle. Intrinsic angular momentum is quantized independently of orbital angular momentum. Additionally, the direction of the spin is also quantized. It has been found that the **magnitude of the intrinsic (internal) spin angular momentum**, S, of an electron is given by
Equation:

$$S = \sqrt{s(s+1)}\frac{h}{2\pi} \quad (s = 1/2 \text{ for electrons}),$$

where s is defined to be the **spin quantum number**. This is very similar to the quantization of L given in $L = \sqrt{l(l+1)}\frac{h}{2\pi}$, except that the only value allowed for s for electrons is 1/2.

The *direction of intrinsic spin is quantized*, just as is the direction of orbital angular momentum. The direction of spin angular momentum along one direction in space, again called the z-axis, can have only the values

Equation:

$$S_z = m_s \frac{h}{2\pi} \quad \left(m_s = -\frac{1}{2}, + \frac{1}{2} \right)$$

for electrons. S_z is the z**-component of spin angular momentum** and m_s is the **spin projection quantum number**. For electrons, s can only be 1/2, and m_s can be either +1/2 or –1/2. Spin projection m_s=$+1/2$ is referred to as *spin up*, whereas $m_s = -1/2$ is called *spin down*. These are illustrated in [link].

Note:
Intrinsic Spin
In later chapters, we will see that intrinsic spin is a characteristic of all subatomic particles. For some particles s is half-integral, whereas for others s is integral—there are crucial differences between half-integral spin particles and integral spin particles. Protons and neutrons, like electrons, have $s = 1/2$, whereas photons have $s = 1$, and other particles called pions have $s = 0$, and so on.

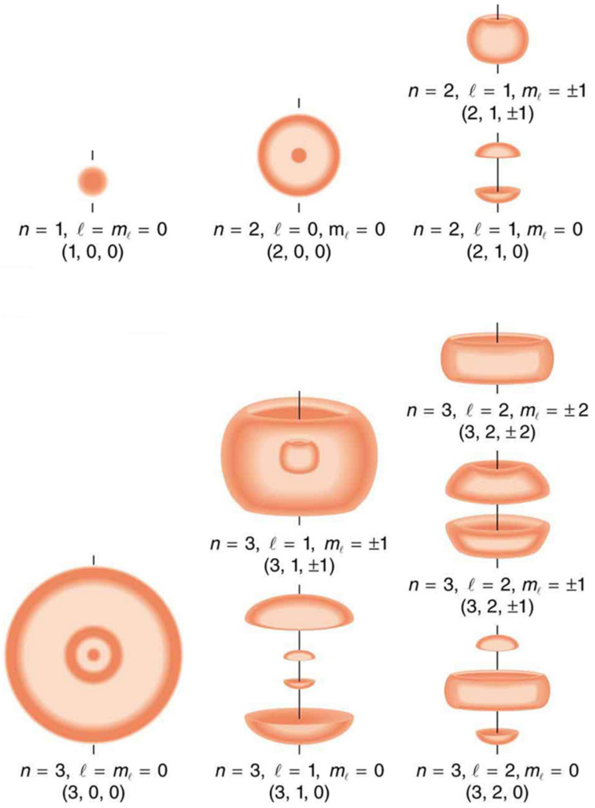
To summarize, the state of a system, such as the precise nature of an electron in an atom, is determined by its particular quantum numbers. These are expressed in the form $(n,\, l,\, m_l,\, m_s)$ —see [link] *For electrons in atoms*, the principal quantum number can have the values $n = 1,\, 2,\, 3,\, ...$. Once n is known, the values of the angular momentum quantum number are limited to $l = 1,\, 2,\, 3,\, ...,n-1$. For a given value of l, the angular momentum projection quantum number can have only the values $m_l = -l,\; -l\; +1,\, ...,\, -1,\, 0,\, 1,\, ...,\, l-1,\, l$. Electron spin is independent of $n,\, l$, and m_l, always having $s = 1/2$. The spin projection quantum number can have two values, $m_s = 1/2$ or $\; -1/2$.

Name	Symbol	Allowed values
Principal quantum number	n	$1,\, 2,\, 3,\, ...$
Angular momentum	l	$0,\, 1,\, 2,\, ...n-1$
Angular momentum projection	m_l	$-l,\; -l\; +1,\, ...,\; -1,\, 0,\, 1,\, ...,\, l-1,\, l$ (or $0,\, \pm 1,\, \pm 2,\, ...,\; \pm\, l$)

Name	Symbol	Allowed values
Spin[footnote] The spin quantum number *s* is usually not stated, since it is always 1/2 for electrons	s	$1/2(\text{electrons})$
Spin projection	m_s	$-1/2,\ +1/2$

Atomic Quantum Numbers

[link] shows several hydrogen states corresponding to different sets of quantum numbers. Note that these clouds of probability are the locations of electrons as determined by making repeated measurements—each measurement finds the electron in a definite location, with a greater chance of finding the electron in some places rather than others. With repeated measurements, the pattern of probability shown in the figure emerges. The clouds of probability do not look like nor do they correspond to classical orbits. The uncertainty principle actually prevents us and nature from knowing how the electron gets from one place to another, and so an orbit really does not exist as such. Nature on a small scale is again much different from that on the large scale.


$n = 2,\ \ell = 1,\ m_\ell = \pm 1$
(2, 1, ±1)
$n = 1,\ \ell = m_\ell = 0$
(1, 0, 0)
$n = 2,\ \ell = 0,\ m_\ell = 0$
(2, 0, 0)
$n = 2,\ \ell = 1,\ m_\ell = 0$
(2, 1, 0)
$n = 3,\ \ell = 2,\ m_\ell = \pm 2$
(3, 2, ±2)
$n = 3,\ \ell = 1,\ m_\ell = \pm 1$
(3, 1, ±1)
$n = 3,\ \ell = 2,\ m_\ell = \pm 1$
(3, 2, ±1)
$n = 3,\ \ell = m_\ell = 0$
(3, 0, 0)
$n = 3,\ \ell = 1,\ m_\ell = 0$
(3, 1, 0)
$n = 3,\ \ell = 2,\ m_\ell = 0$
(3, 2, 0)

Probability clouds for the electron in the ground state and several excited states of hydrogen. The nature of these states is determined by their sets of quantum numbers, here given as (n, l, m_l). The ground state is (0, 0, 0); one of the possibilities for the second excited state is (3, 2, 1). The probability of finding the electron is indicated by the shade of color; the darker the coloring the greater the chance of finding the electron.

We will see that the quantum numbers discussed in this section are valid for a broad range of particles and other systems, such as nuclei. Some quantum numbers, such as intrinsic spin, are related to fundamental classifications of subatomic particles, and they obey laws that will give us further insight into the substructure of matter and its interactions.

Note:
PhET Explorations: Stern-Gerlach Experiment
The classic Stern-Gerlach Experiment shows that atoms have a property called spin. Spin is a kind of intrinsic angular momentum, which has no classical counterpart. When the z-component of the spin is

measured, one always gets one of two values: spin up or spin down.

https://phet.colorado.edu/sims/stern-gerlach/stern-gerlach_en.html

Section Summary

- Quantum numbers are used to express the allowed values of quantized entities. The principal quantum number n labels the basic states of a system and is given by
  Equation:
  $$n = 1, 2, 3, \ldots.$$
- The magnitude of angular momentum is given by
  Equation:
  $$L = \sqrt{l(l+1)}\frac{h}{2\pi} \quad (l = 0, 1, 2, \ldots, n-1),$$
  where l is the angular momentum quantum number. The direction of angular momentum is quantized, in that its component along an axis defined by a magnetic field, called the z-axis is given by
  Equation:
  $$L_z = m_l\frac{h}{2\pi} \quad (m_l = -l, -l+1, \ldots, -1, 0, 1, \ldots l-1, l),$$
  where L_z is the z-component of the angular momentum and m_l is the angular momentum projection quantum number. Similarly, the electron's intrinsic spin angular momentum S is given by
  Equation:
  $$S = \sqrt{s(s+1)}\frac{h}{2\pi} \quad (s = 1/2 \text{ for electrons}),$$
  s is defined to be the spin quantum number. Finally, the direction of the electron's spin along the z-axis is given by
  Equation:
  $$S_z = m_s\frac{h}{2\pi} \quad \left(m_s = -\frac{1}{2}, +\frac{1}{2}\right),$$
  where S_z is the z-component of spin angular momentum and m_s is the spin projection quantum number. Spin projection m_s=$+1/2$ is referred to as spin up, whereas $m_s = -1/2$ is called spin down. [link] summarizes the atomic quantum numbers and their allowed values.

Conceptual Questions

Exercise:

Problem: Define the quantum numbers n, l, m_l, s, and m_s.

Exercise:

Problem: For a given value of n, what are the allowed values of l?

Exercise:

Problem:

For a given value of l, what are the allowed values of m_l? What are the allowed values of m_l for a given value of n? Give an example in each case.

Exercise:

Problem:

List all the possible values of s and m_s for an electron. Are there particles for which these values are different? The same?

Problem Exercises

Exercise:

Problem:

If an atom has an electron in the $n = 5$ state with $m_l = 3$, what are the possible values of l?

Solution:

$l = 4, 3$ are possible since $l < n$ and $\mid m_l \mid \leq l$.

Exercise:

Problem: An atom has an electron with $m_l = 2$. What is the smallest value of n for this electron?

Exercise:

Problem: What are the possible values of m_l for an electron in the $n = 4$ state?

Solution:

$n = 4 \Rightarrow l = 3, 2, 1, 0 \Rightarrow m_l = \pm 3, \ \pm 2, \ \pm 1, 0$ are possible.

Exercise:

Problem:

What, if any, constraints does a value of $m_l = 1$ place on the other quantum numbers for an electron in an atom?

Exercise:

Problem:

(a) Calculate the magnitude of the angular momentum for an $l=1$ electron. (b) Compare your answer to the value Bohr proposed for the $n=1$ state.

Solution:

(a) $1.49\times10^{-34}\ \mathrm{J\cdot s}$

(b) $1.06\times10^{-34}\ \mathrm{J\cdot s}$

Exercise:

Problem:

(a) What is the magnitude of the angular momentum for an $l=1$ electron? (b) Calculate the magnitude of the electron’s spin angular momentum. (c) What is the ratio of these angular momenta?

Exercise:

Problem: Repeat [link] for $l=3$.

Solution:

(a) $3.66\times10^{-34}\ \mathrm{J\cdot s}$

(b) $s=9.13\times10^{-35}\ \mathrm{J\cdot s}$

(c) $\frac{L}{S}=\frac{\sqrt{12}}{\sqrt{3/4}}=4$

Exercise:

Problem:

(a) How many angles can L make with the z-axis for an $l=2$ electron? (b) Calculate the value of the smallest angle.

Exercise:

Problem: What angles can the spin S of an electron make with the z-axis?

Solution:

$\theta=54.7^\mathrm{o}, 125.3^\mathrm{o}$

Glossary

quantum numbers
: the values of quantized entities, such as energy and angular momentum

angular momentum quantum number

a quantum number associated with the angular momentum of electrons

spin quantum number
the quantum number that parameterizes the intrinsic angular momentum (or spin angular momentum, or simply spin) of a given particle

spin projection quantum number
quantum number that can be used to calculate the intrinsic electron angular momentum along the z -axis

z-component of spin angular momentum
component of intrinsic electron spin along the z-axis

magnitude of the intrinsic (internal) spin angular momentum
given by $S = \sqrt{s(s+1)}\frac{h}{2\pi}$

z-component of the angular momentum
component of orbital angular momentum of electron along the z-axis

The Pauli Exclusion Principle

- Define the composition of an atom along with its electrons, neutrons, and protons.
- Explain the Pauli exclusion principle and its application to the atom.
- Specify the shell and subshell symbols and their positions.
- Define the position of electrons in different shells of an atom.
- State the position of each element in the periodic table according to shell filling.

Multiple-Electron Atoms

All atoms except hydrogen are multiple-electron atoms. The physical and chemical properties of elements are directly related to the number of electrons a neutral atom has. The periodic table of the elements groups elements with similar properties into columns. This systematic organization is related to the number of electrons in a neutral atom, called the **atomic number**, Z. We shall see in this section that the exclusion principle is key to the underlying explanations, and that it applies far beyond the realm of atomic physics.

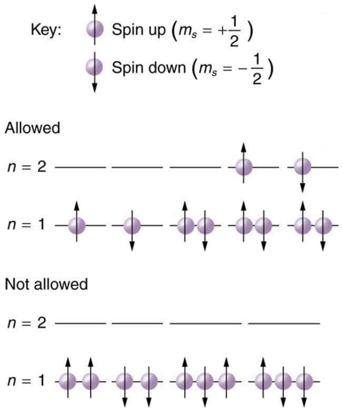
In 1925, the Austrian physicist Wolfgang Pauli (see [link]) proposed the following rule: No two electrons can have the same set of quantum numbers. That is, no two electrons can be in the same state. This statement is known as the **Pauli exclusion principle**, because it excludes electrons from being in the same state. The Pauli exclusion principle is extremely powerful and very broadly applicable. It applies to any identical particles with half-integral intrinsic spin—that is, having $s = 1/2, 3/2, ...$ Thus no two electrons can have the same set of quantum numbers.

Note:
Pauli Exclusion Principle
No two electrons can have the same set of quantum numbers. That is, no two electrons can be in the same state.

The Austrian physicist Wolfgang Pauli (1900–1958) played a major role in the development of quantum mechanics. He proposed the exclusion principle; hypothesized the existence of an important particle, called the neutrino, before it was directly observed; made fundamental contributions to several areas of theoretical physics; and influenced many students who went on to do important work of their own. (credit: Nobel Foundation, via Wikimedia Commons)

Let us examine how the exclusion principle applies to electrons in atoms. The quantum numbers involved were defined in Quantum Numbers and Rules as n, l, m_l, s, and m_s. Since s is always $1/2$ for electrons, it is redundant to list s, and so we omit it and specify the state of an electron by a set of four numbers $(n,\,l,\,m_l,\,m_s)$. For example, the quantum numbers $(2,\,1,\,0,\,-1/2)$ completely specify the state of an electron in an atom.

Since no two electrons can have the same set of quantum numbers, there are limits to how many of them can be in the same energy state. Note that n determines the energy state in the absence of a magnetic field. So we first choose n, and then we see how many electrons can be in this energy state or energy level. Consider the $n=1$ level, for example. The only value l can have is 0 (see [link] for a list of possible values once n is known), and thus m_l can only be 0. The spin projection m_s can be either $+1/2$ or $-1/2$, and so there can be two electrons in the $n=1$ state. One has quantum numbers $(1,\,0,\,0,\,+1/2)$, and the other has $(1,\,0,\,0,\,-1/2)$. [link] illustrates that there can be one or two electrons having $n=1$, but not three.


Key:
Spin up ($m_s = +\frac{1}{2}$)
Spin down ($m_s = -\frac{1}{2}$)
Allowed
n = 2
n = 1
Not allowed
n = 2
n = 1

The Pauli exclusion principle explains why some configurations of electrons are allowed while others are not. Since electrons cannot have the same set of quantum numbers, a maximum of two can be in the $n = 1$ level, and a third electron must reside in the higher-energy $n = 2$ level. If there are two electrons in the $n = 1$ level, their spins must be in opposite directions. (More precisely, their spin projections must differ.)

Shells and Subshells

Because of the Pauli exclusion principle, only hydrogen and helium can have all of their electrons in the $n = 1$ state. Lithium (see the periodic table) has three electrons, and so one must be in the $n = 2$ level. This leads to the concept of shells and shell filling. As we progress up in the number of electrons, we go from hydrogen to helium, lithium, beryllium, boron, and so on, and we see that there are limits to the number of electrons for each value of n . Higher values of the shell n correspond to higher energies, and they can allow more electrons because of the various combinations of l, m_l, and m_s that are possible. Each value of the principal quantum number n thus corresponds to an atomic **shell** into which a limited number of electrons can go. Shells and the number of electrons in them determine the physical and chemical properties of atoms, since it is the outermost electrons that interact most with anything outside the atom.

The probability clouds of electrons with the lowest value of l are closest to the nucleus and, thus, more tightly bound. Thus when shells fill, they start with $l = 0$, progress to $l = 1$, and so on. Each value of l thus corresponds to a **subshell**.

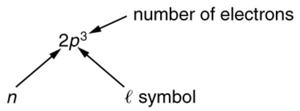
The table given below lists symbols traditionally used to denote shells and subshells.

Shell	**Subshell**	
n	l	***Symbol***
1	0	s
2	1	p
3	2	d
4	3	f

Shell	Subshell	
5	4	g
	5	h
	6[footnote] It is unusual to deal with subshells having l greater than 6, but when encountered, they continue to be labeled in alphabetical order.	i

Shell and Subshell Symbols

To denote shells and subshells, we write nl with a number for n and a letter for l. For example, an electron in the $n=1$ state must have $l=0$, and it is denoted as a $1s$ electron. Two electrons in the $n=1$ state is denoted as $1s^2$. Another example is an electron in the $n=2$ state with $l=1$, written as $2p$. The case of three electrons with these quantum numbers is written $2p^3$. This notation, called spectroscopic notation, is generalized as shown in [link].



Counting the number of possible combinations of quantum numbers allowed by the exclusion principle, we can determine how many electrons it takes to fill each subshell and shell.

Example:
How Many Electrons Can Be in This Shell?

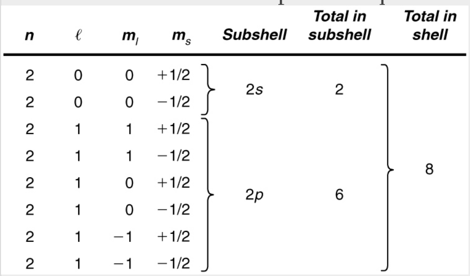
List all the possible sets of quantum numbers for the $n=2$ shell, and determine the number of electrons that can be in the shell and each of its subshells.

Strategy

Given $n=2$ for the shell, the rules for quantum numbers limit l to be 0 or 1. The shell therefore has two subshells, labeled $2s$ and $2p$. Since the lowest l subshell fills first, we start with the $2s$ subshell possibilities and then proceed with the $2p$ subshell.

Solution

It is convenient to list the possible quantum numbers in a table, as shown below.



n	ℓ	m_l	m_s	Subshell	Total in subshell	Total in shell
2	0	0	+1/2	$2s$	2	8
2	0	0	−1/2			
2	1	1	+1/2	$2p$	6	
2	1	1	−1/2			
2	1	0	+1/2			
2	1	0	−1/2			
2	1	−1	+1/2			
2	1	−1	−1/2			

Discussion

It is laborious to make a table like this every time we want to know how many electrons can be in a shell or subshell. There exist general rules that are easy to apply, as we shall now see.

The number of electrons that can be in a subshell depends entirely on the value of l. Once l is known, there are a fixed number of values of m_l, each of which can have two values for m_s First, since m_l goes from $-l$ to l in steps of 1, there are $2l+1$ possibilities. This number is multiplied by 2, since each electron can be spin up or spin down. Thus the *maximum number of electrons that can be in a subshell* is $2(2l+1)$.

For example, the $2s$ subshell in [link] has a maximum of 2 electrons in it, since $2(2l+1)=2(0+1)=2$ for this subshell. Similarly, the $2p$ subshell has a maximum of 6 electrons, since $2(2l+1)=2(2+1)=6$. For a shell, the maximum number is the sum of what can fit in the subshells. Some algebra shows that the *maximum number of electrons that can be in a shell* is $2n^2$.

For example, for the first shell $n=1$, and so $2n^2=2$. We have already seen that only two electrons can be in the $n=1$ shell. Similarly, for the second shell, $n=2$, and so $2n^2=8$. As found in [link], the total number of electrons in the $n=2$ shell is 8.

Example:
Subshells and Totals for $n=3$
How many subshells are in the $n=3$ shell? Identify each subshell, calculate the maximum number of electrons that will fit into each, and verify that the total is $2n^2$.
Strategy
Subshells are determined by the value of l; thus, we first determine which values of l are allowed, and then we apply the equation "maximum number of electrons that can be in a subshell $=2(2l+1)$" to find the number of electrons in each subshell.
Solution
Since $n=3$, we know that l can be 0, 1, or 2; thus, there are three possible subshells. In standard notation, they are labeled the $3s$, $3p$, and $3d$ subshells. We have already seen that 2 electrons can be in an s state, and 6 in a p state, but let us use the equation "maximum number of electrons that can be in a subshell = $2(2l+1)$" to calculate the maximum number in each:
Equation:

$$\begin{array}{l}3s\text{ has }l=0\text{; thus, }2(2l+1)=2(0+1)=2\\ 3p\text{ has }l=1\text{; thus, }2(2l+1)=2(2+1)=6\\ 3d\text{ has }l=2\text{; thus, }2(2l+1)=2(4+1)=10\\ \text{Total}=18\\ (\text{in the }n=3\text{ shell})\end{array}$$

The equation "maximum number of electrons that can be in a shell = $2n^2$" gives the maximum number in the $n=3$ shell to be
Equation:

$$\text{Maximum number of electrons}=2n^2=2(3)^2=2(9)=18.$$

Discussion
The total number of electrons in the three possible subshells is thus the same as the formula $2n^2$. In standard (spectroscopic) notation, a filled $n=3$ shell is denoted as $3s^2 3p^6 3d^{10}$. Shells do not fill in a simple manner. Before the $n=3$ shell is completely filled, for example, we begin to find electrons in the $n=4$ shell.

Shell Filling and the Periodic Table

[link] shows electron configurations for the first 20 elements in the periodic table, starting with hydrogen and its single electron and ending with calcium. The Pauli exclusion principle determines the maximum number of electrons allowed in each shell and subshell. But the order in which the shells and subshells are filled is complicated because of the large numbers of interactions between electrons.

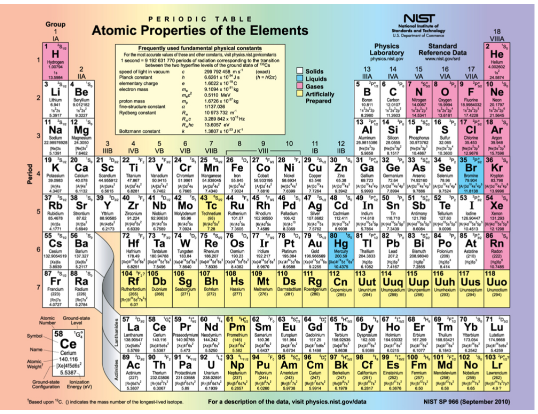
Element	**Number of electrons (Z)**	**Ground state configuration**					
H	1	$1s^1$					
He	2	$1s^2$					
Li	3	$1s^2$	$2s^1$				
Be	4	"	$2s^2$				
B	5	"	$2s^2$	$2p^1$			
C	6	"	$2s^2$	$2p^2$			
N	7	"	$2s^2$	$2p^3$			
O	8	"	$2s^2$	$2p^4$			
F	9	"	$2s^2$	$2p^5$			
Ne	10	"	$2s^2$	$2p^6$			
Na	11	"	$2s^2$	$2p^6$	$3s^1$		
Mg	12	"	"	"	$3s^2$		
Al	13	"	"	"	$3s^2$	$3p^1$	
Si	14	"	"	"	$3s^2$	$3p^2$	

Element	Number of electrons (Z)	Ground state configuration					
P	15	"	"	"	$3s^2$	$3p^3$	
S	16	"	"	"	$3s^2$	$3p^4$	
Cl	17	"	"	"	$3s^2$	$3p^5$	
Ar	18	"	"	"	$3s^2$	$3p^6$	
K	19	"	"	"	$3s^2$	$3p^6$	$4s^1$
Ca	20	"	"	"	"	"	$4s^2$

Electron Configurations of Elements Hydrogen Through Calcium

Examining the above table, you can see that as the number of electrons in an atom increases from 1 in hydrogen to 2 in helium and so on, the lowest-energy shell gets filled first—that is, the $n = 1$ shell fills first, and then the $n = 2$ shell begins to fill. Within a shell, the subshells fill starting with the lowest l, or with the s subshell, then the p, and so on, usually until all subshells are filled. The first exception to this occurs for potassium, where the $4s$ subshell begins to fill before any electrons go into the $3d$ subshell. The next exception is not shown in [link]; it occurs for rubidium, where the $5s$ subshell starts to fill before the $4d$ subshell. The reason for these exceptions is that $l = 0$ electrons have probability clouds that penetrate closer to the nucleus and, thus, are more tightly bound (lower in energy).

[link] shows the periodic table of the elements, through element 118. Of special interest are elements in the main groups, namely, those in the columns numbered 1, 2, 13, 14, 15, 16, 17, and 18.


PERIODIC TABLE
Atomic Properties of the Elements
NIST
Physics Laboratory
Standard Reference Data
Frequently used fundamental physical constants
Solids
Liquids
Gases
Artificially Prepared
Group
Period
Lanthanides
Actinides
For a description of the data, visit physics.nist.gov/data
NIST SP 966 (September 2010)

Periodic table of the elements (credit:

National Institute of Standards and Technology, U.S. Department of Commerce)

The number of electrons in the outermost subshell determines the atom’s chemical properties, since it is these electrons that are farthest from the nucleus and thus interact most with other atoms. If the outermost subshell can accept or give up an electron easily, then the atom will be highly reactive chemically. Each group in the periodic table is characterized by its outermost electron configuration. Perhaps the most familiar is Group 18 (Group VIII), the noble gases (helium, neon, argon, etc.). These gases are all characterized by a filled outer subshell that is particularly stable. This means that they have large ionization energies and do not readily give up an electron. Furthermore, if they were to accept an extra electron, it would be in a significantly higher level and thus loosely bound. Chemical reactions often involve sharing electrons. Noble gases can be forced into unstable chemical compounds only under high pressure and temperature.

Group 17 (Group VII) contains the halogens, such as fluorine, chlorine, iodine and bromine, each of which has one less electron than a neighboring noble gas. Each halogen has 5 p electrons (a p^5 configuration), while the p subshell can hold 6 electrons. This means the halogens have one vacancy in their outermost subshell. They thus readily accept an extra electron (it becomes tightly bound, closing the shell as in noble gases) and are highly reactive chemically. The halogens are also likely to form singly negative ions, such as $\mathrm{C1}^-$, fitting an extra electron into the vacancy in the outer subshell. In contrast, alkali metals, such as sodium and potassium, all have a single s electron in their outermost subshell (an s^1 configuration) and are members of Group 1 (Group I). These elements easily give up their extra electron and are thus highly reactive chemically. As you might expect, they also tend to form singly positive ions, such as Na^+, by losing their loosely bound outermost electron. They are metals (conductors), because the loosely bound outer electron can move freely.

Of course, other groups are also of interest. Carbon, silicon, and germanium, for example, have similar chemistries and are in Group 4 (Group IV). Carbon, in particular, is extraordinary in its ability to form many types of bonds and to be part of long chains, such as inorganic molecules. The large group of what are called transitional elements is characterized by the filling of the d subshells and crossing of energy levels. Heavier groups, such as the lanthanide series, are more complex—their shells do not fill in simple order. But the groups recognized by chemists such as Mendeleev have an explanation in the substructure of atoms.

Note:
PhET Explorations: Stern-Gerlach Experiment
Build an atom out of protons, neutrons, and electrons, and see how the element, charge, and mass change. Then play a game to test your ideas!

https://phet.colorado.edu/sims/html/build-an-atom/latest/build-an-atom_en.html

Section Summary

- The state of a system is completely described by a complete set of quantum numbers. This set is written as (n, l, m_l, m_s).
- The Pauli exclusion principle says that no two electrons can have the same set of quantum numbers; that is, no two electrons can be in the same state.
- This exclusion limits the number of electrons in atomic shells and subshells. Each value of n corresponds to a shell, and each value of l corresponds to a subshell.
- The maximum number of electrons that can be in a subshell is $2(2l+1)$.
- The maximum number of electrons that can be in a shell is $2n^2$.

Conceptual Questions

Exercise:

Problem:

Identify the shell, subshell, and number of electrons for the following: (a) $2p^3$. (b) $4d^9$. (c) $3s^1$. (d) $5g^{16}$.

Exercise:

Problem:

Which of the following are not allowed? State which rule is violated for any that are not allowed. (a) $1p^3$ (b) $2p^8$(c) $3g^{11}$ (d) $4f^2$

Problem Exercises

Exercise:

Problem: (a) How many electrons can be in the $n = 4$ shell?

(b) What are its subshells, and how many electrons can be in each?

Solution:

(a) 32. (b) $2 \text{ in } s$, $6 \text{ in } p$, $10 \text{ in } d$, and $14 \text{ in } f$, for a total of 32.

Exercise:

Problem: (a) What is the minimum value of 1 for a subshell that has 11 electrons in it?

(b) If this subshell is in the $n = 5$ shell, what is the spectroscopic notation for this atom?

Exercise:

Problem:

(a) If one subshell of an atom has 9 electrons in it, what is the minimum value of l? (b) What is the spectroscopic notation for this atom, if this subshell is part of the $n = 3$ shell?

Solution:

(a) 2

(b) $3d^9$

Exercise:

Problem:

(a) List all possible sets of quantum numbers (n, l, m_l, m_s) for the $n = 3$ shell, and determine the number of electrons that can be in the shell and each of its subshells.

(b) Show that the number of electrons in the shell equals $2n^2$ and that the number in each subshell is $2(2l + 1)$.

Exercise:

Problem:

Which of the following spectroscopic notations are not allowed? (a) $5s^1$ (b) $1d^1$ (c) $4s^3$ (d) $3p^7$ (e) $5g^{15}$. State which rule is violated for each that is not allowed.

Solution:

(b) $n \geq l$ is violated,

(c) cannot have 3 electrons in s subshell since $3 > (2l + 1) = 2$

(d) cannot have 7 electrons in p subshell since $7 > (2l + 1) = 2(2 + 1) = 6$

Exercise:

Problem:

Which of the following spectroscopic notations are allowed (that is, which violate none of the rules regarding values of quantum numbers)? (a) $1s^1$(b) $1d^3$(c) $4s^2$ (d) $3p^7$(e) $6h^{20}$

Exercise:

Problem:

(a) Using the Pauli exclusion principle and the rules relating the allowed values of the quantum numbers (n, l, m_l, m_s), prove that the maximum number of electrons in a subshell is $2n^2$.

(b) In a similar manner, prove that the maximum number of electrons in a shell is $2n^2$.

Solution:

(a) The number of different values of m_l is $\pm l, \pm (l - 1), ...,0$ for each $l > 0$ and one for $l = 0 \Rightarrow (2l + 1)$. Also an overall factor of 2 since each m_l can have m_s equal to either $+1/2$ or $-1/2 \Rightarrow 2(2l + 1)$.

(b) for each value of l, you get $2(2l + 1)$

$= 0, 1, 2, ...,(n–1) \Rightarrow 2\{[(2)(0) + 1] + [(2)(1) + 1] + ... + [(2)(n - 1) + 1]\} = 2[1 + 3 + ... + (2n - 3) +$

n terms

to see that the expression in the box is $= n^2$, imagine taking $(n - 1)$ from the last term and adding it to first term $= 2[1 + (n–1) + 3 + ... + (2n - 3) + (2n - 1)–(n - 1)] = 2[n + 3 + ... + (2n - 3) + n]$. Now take $(n - 3)$ from penultimate term and add to the second term $2[n + n + ... + n + n] = 2n^2$.

n terms

Exercise:

Problem:Integrated Concepts

Estimate the density of a nucleus by calculating the density of a proton, taking it to be a sphere 1.2 fm in diameter. Compare your result with the value estimated in this chapter.

Exercise:

Problem: Integrated Concepts

The electric and magnetic forces on an electron in the CRT in [link] are supposed to be in opposite directions. Verify this by determining the direction of each force for the situation shown. Explain how you obtain the directions (that is, identify the rules used).

Solution:

The electric force on the electron is up (toward the positively charged plate). The magnetic force is down (by the RHR).

Exercise:

Problem:

(a) What is the distance between the slits of a diffraction grating that produces a first-order maximum for the first Balmer line at an angle of 20.0^{o}?

(b) At what angle will the fourth line of the Balmer series appear in first order?

(c) At what angle will the second-order maximum be for the first line?

Exercise:

Problem: **Integrated Concepts**

A galaxy moving away from the earth has a speed of $0.0100c$. What wavelength do we observe for an $n_{\text{i}} = 7$ to $n_{\text{f}} = 2$ transition for hydrogen in that galaxy?

Solution:

401 nm

Exercise:

Problem: **Integrated Concepts**

Calculate the velocity of a star moving relative to the earth if you observe a wavelength of 91.0 nm for ionized hydrogen capturing an electron directly into the lowest orbital (that is, a $n_{\text{i}} = \infty$ to $n_{\text{f}} = 1$, or a Lyman series transition).

Exercise:

Problem: **Integrated Concepts**

In a Millikan oil-drop experiment using a setup like that in [link], a 500-V potential difference is applied to plates separated by 2.50 cm. (a) What is the mass of an oil drop having two extra electrons that is suspended motionless by the field between the plates? (b) What is the diameter of the drop, assuming it is a sphere with the density of olive oil?

Solution:

(a) $6.54 \times 10^{-16}\ \text{kg}$

(b) $5.54 \times 10^{-7}\ \text{m}$

Exercise:

Problem: **Integrated Concepts**

What double-slit separation would produce a first-order maximum at 3.00^{o} for 25.0-keV x rays? The small answer indicates that the wave character of x rays is best determined by having them interact with very small objects such as atoms and molecules.

Exercise:

Problem: **Integrated Concepts**

In a laboratory experiment designed to duplicate Thomson's determination of q_e/m_e, a beam of electrons having a velocity of $6.00 \times 10^{7}\ \text{m/s}$ enters a $5.00 \times 10^{-3}\ \text{T}$ magnetic field. The beam moves perpendicular to the field in a path having a 6.80-cm radius of curvature. Determine q_e/m_e from these observations, and compare the result with the known value.

Solution:

$1.76 \times 10^{11}\ \text{C/kg}$, which agrees with the known value of $1.759 \times 10^{11}\ \text{C/kg}$ to within the precision of the measurement

Exercise:

Problem:Integrated Concepts

Find the value of l, the orbital angular momentum quantum number, for the moon around the earth. The extremely large value obtained implies that it is impossible to tell the difference between adjacent quantized orbits for macroscopic objects.

Exercise:

Problem: Integrated Concepts

Particles called muons exist in cosmic rays and can be created in particle accelerators. Muons are very similar to electrons, having the same charge and spin, but they have a mass 207 times greater. When muons are captured by an atom, they orbit just like an electron but with a smaller radius, since the mass in $a_{\text{B}} = \frac{h^2}{4\pi^2 m_e k q_e^2} = 0.529 \times 10^{-10}\ \text{m}$ is 207 m_e.

(a) Calculate the radius of the $n = 1$ orbit for a muon in a uranium ion ($Z = 92$).

(b) Compare this with the 7.5-fm radius of a uranium nucleus. Note that since the muon orbits inside the electron, it falls into a hydrogen-like orbit. Since your answer is less than the radius of the nucleus, you can see that the photons emitted as the muon falls into its lowest orbit can give information about the nucleus.

Solution:

(a) 2.78 fm

(b) 0.37 of the nuclear radius.

Exercise:

Problem:Integrated Concepts

Calculate the minimum amount of energy in joules needed to create a population inversion in a helium-neon laser containing 1.00×10^{-4} moles of neon.

Exercise:

Problem: Integrated Concepts

A carbon dioxide laser used in surgery emits infrared radiation with a wavelength of $10.6\ \mu\text{m}$. In 1.00 ms, this laser raised the temperature of $1.00\ \text{cm}^3$ of flesh to 100ºC and evaporated it.

(a) How many photons were required? You may assume flesh has the same heat of vaporization as water. (b) What was the minimum power output during the flash?

Solution:

(a) 1.34×10^{23}

(b) 2.52 MW

Exercise:

Problem: **Integrated Concepts**

Suppose an MRI scanner uses 100-MHz radio waves.

(a) Calculate the photon energy.

(b) How does this compare to typical molecular binding energies?

Exercise:

Problem: **Integrated Concepts**

(a) An excimer laser used for vision correction emits 193-nm UV. Calculate the photon energy in eV.

(b) These photons are used to evaporate corneal tissue, which is very similar to water in its properties. Calculate the amount of energy needed per molecule of water to make the phase change from liquid to gas. That is, divide the heat of vaporization in kJ/kg by the number of water molecules in a kilogram.

(c) Convert this to eV and compare to the photon energy. Discuss the implications.

Solution:

(a) 6.42 eV

(b) $7.27 \times 10^{-20}\ \text{J/molecule}$

(c) 0.454 eV, 14.1 times less than a single UV photon. Therefore, each photon will evaporate approximately 14 molecules of tissue. This gives the surgeon a rather precise method of removing corneal tissue from the surface of the eye.

Exercise:

Problem: **Integrated Concepts**

A neighboring galaxy rotates on its axis so that stars on one side move toward us as fast as 200 km/s, while those on the other side move away as fast as 200 km/s. This causes the EM radiation we receive to be Doppler shifted by velocities over the entire range of ±200 km/s. What range of wavelengths will we observe for the 656.0-nm line in the Balmer series of hydrogen emitted by stars in this galaxy. (This is called line broadening.)

Exercise:

Problem: **Integrated Concepts**

A pulsar is a rapidly spinning remnant of a supernova. It rotates on its axis, sweeping hydrogen along with it so that hydrogen on one side moves toward us as fast as 50.0 km/s, while that on the other side moves away as fast as 50.0 km/s. This means that the EM radiation we receive will be Doppler shifted over a range of $\pm 50.0\ \text{km/s}$. What range of wavelengths will we observe for the 91.20-nm line in the Lyman series of hydrogen? (Such line broadening is observed and actually provides part of the evidence for rapid rotation.)

Solution:

91.18 nm to 91.22 nm

Exercise:

Problem: **Integrated Concepts**

Prove that the velocity of charged particles moving along a straight path through perpendicular electric and magnetic fields is $v = E/B$. Thus crossed electric and magnetic fields can be used as a velocity selector independent of the charge and mass of the particle involved.

Exercise:

Problem: **Unreasonable Results**

(a) What voltage must be applied to an X-ray tube to obtain 0.0100-fm-wavelength X-rays for use in exploring the details of nuclei? (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Solution:

(a) $1.24 \times 10^{11}\text{ V}$

(b) The voltage is extremely large compared with any practical value.

(c) The assumption of such a short wavelength by this method is unreasonable.

Exercise:

Problem: **Unreasonable Results**

A student in a physics laboratory observes a hydrogen spectrum with a diffraction grating for the purpose of measuring the wavelengths of the emitted radiation. In the spectrum, she observes a yellow line and finds its wavelength to be 589 nm. (a) Assuming this is part of the Balmer series, determine n_{i}, the principal quantum number of the initial state. (b) What is unreasonable about this result? (c) Which assumptions are unreasonable or inconsistent?

Exercise:

Problem: **Construct Your Own Problem**

The solar corona is so hot that most atoms in it are ionized. Consider a hydrogen-like atom in the corona that has only a single electron. Construct a problem in which you calculate selected spectral energies and wavelengths of the Lyman, Balmer, or other series of this atom that could be used to identify its presence in a very hot gas. You will need to choose the atomic number of the atom, identify the element, and choose which spectral lines to consider.

Exercise:

Problem: **Construct Your Own Problem**

Consider the Doppler-shifted hydrogen spectrum received from a rapidly receding galaxy. Construct a problem in which you calculate the energies of selected spectral lines in the Balmer series and examine whether they can be described with a formula like that in the equation $\frac{1}{\lambda} = R \ \ \frac{1}{n_{\text{f}}^2} - \frac{1}{n_{\text{i}}^2}$, but with a different constant R.

Glossary

atomic number
: the number of protons in the nucleus of an atom

Pauli exclusion principle
a principle that states that no two electrons can have the same set of quantum numbers; that is, no two electrons can be in the same state

shell
a probability cloud for electrons that has a single principal quantum number

subshell
the probability cloud for electrons that has a single angular momentum quantum number l

Introduction to Frontiers of Physics class="introduction"

This galaxy is ejecting huge jets of matter, powered by an immensely massive black hole at its center. (credit: X-ray: NASA/CXC/CfA/R. Kraft et al.)

Frontiers are exciting. There is mystery, surprise, adventure, and discovery. The satisfaction of finding the answer to a question is made keener by the fact that the answer always leads to a new question. The picture of nature becomes more complete, yet nature retains its sense of mystery and never loses its ability to awe us. The view of physics is beautiful looking both backward and forward in time. What marvelous patterns we have discovered. How clever nature seems in its rules and connections. How awesome. And we continue looking ever deeper and ever further, probing the basic structure of matter, energy, space, and time and wondering about the scope of the universe, its beginnings and future.

You are now in a wonderful position to explore the forefronts of physics, both the new discoveries and the unanswered questions. With the concepts, qualitative and quantitative, the problem-solving skills, the feeling for connections among topics, and all the rest you have mastered, you can more deeply appreciate and enjoy the brief treatments that follow. Years from now you will still enjoy the quest with an insight all the greater for your efforts.

Cosmology and Particle Physics

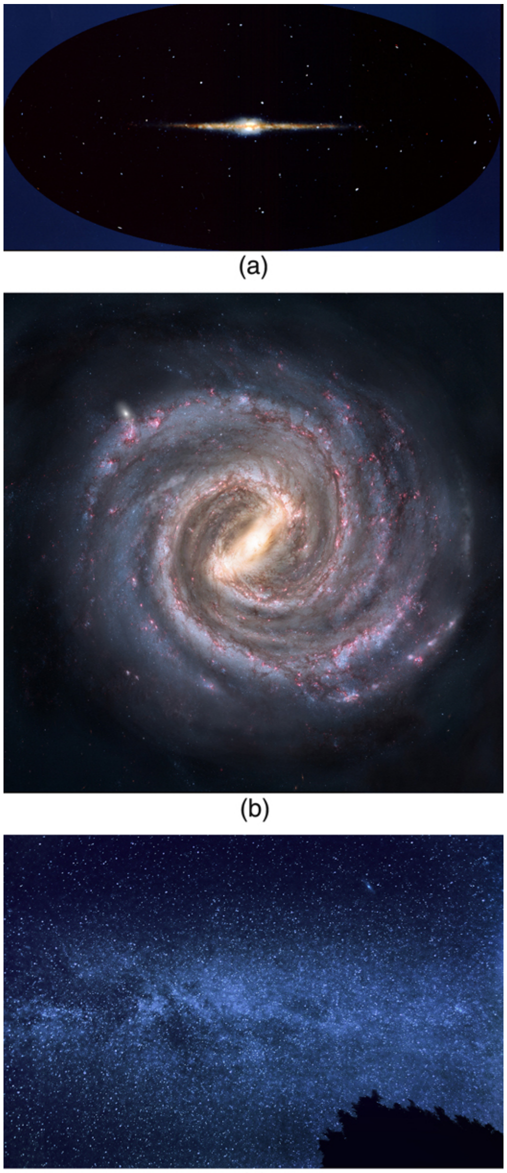
- Discuss the expansion of the universe.
- Explain the Big Bang.

Look at the sky on some clear night when you are away from city lights. There you will see thousands of individual stars and a faint glowing background of millions more. The Milky Way, as it has been called since ancient times, is an arm of our galaxy of stars—the word *galaxy* coming from the Greek word *galaxias*, meaning milky. We know a great deal about our Milky Way galaxy and of the billions of other galaxies beyond its fringes. But they still provoke wonder and awe (see [link]). And there are still many questions to be answered. Most remarkable when we view the universe on the large scale is that once again explanations of its character and evolution are tied to the very small scale. Particle physics and the questions being asked about the very small scales may also have their answers in the very large scales.

Take a moment to contemplate these clusters of galaxies, photographed by the Hubble Space Telescope. Trillions of stars linked by gravity in fantastic forms, glowing with light and showing evidence of undiscovered matter. What are they like, these myriad stars? How did they evolve? What can they tell us of matter, energy, space, and time? (credit: NASA, ESA, K. Sharon (Tel Aviv University) and E. Ofek (Caltech))

As has been noted in numerous Things Great and Small vignettes, this is not the first time the large has been explained by the small and vice versa. Newton realized that the nature of gravity on Earth that pulls an apple to the ground could explain the motion of the moon and planets so much farther away. Minute atoms and molecules explain the chemistry of substances on a much larger scale. Decays of tiny nuclei explain the hot interior of the Earth. Fusion of nuclei likewise explains the energy of stars. Today, the patterns in particle physics seem to be explaining the evolution and character of the universe. And the nature of the universe has implications for unexplored regions of particle physics.

Cosmology is the study of the character and evolution of the universe. What are the major characteristics of the universe as we know them today? First, there are approximately 10^{11} galaxies in the observable part of the universe. An average galaxy contains more than 10^{11} stars, with our Milky Way galaxy being larger than average, both in its number of stars and its dimensions. Ours is a spiral-shaped galaxy with a diameter of about 100,000 light years and a thickness of about 2000 light years in the arms with a central bulge about 10,000 light years across. The Sun lies about 30,000 light years from the center near the galactic plane. There are significant clouds of gas, and there is a halo of less-dense regions of stars surrounding the main body. (See [link].) Evidence strongly suggests the existence of a large amount of additional matter in galaxies that does not produce light—the mysterious dark matter we shall later discuss.



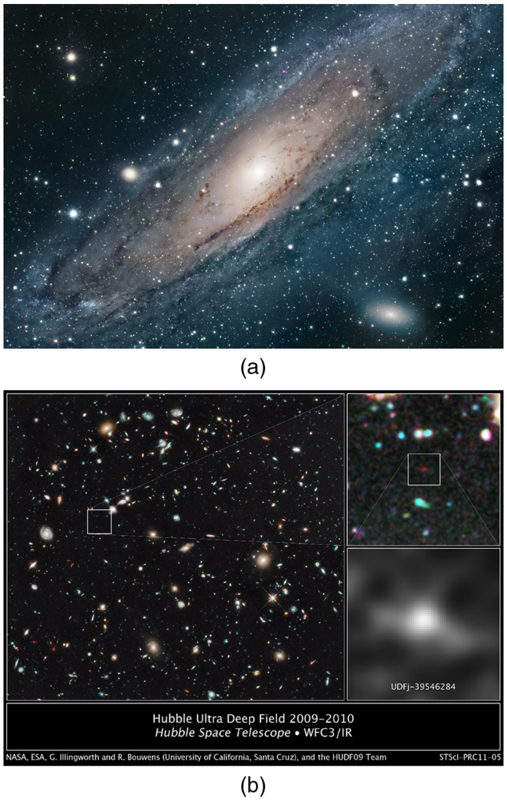
(a)

(b)

(c)

The Milky Way galaxy is typical of large spiral galaxies in its size, its shape, and the presence of gas and dust. We are fortunate to be in a location where we can see out of the galaxy and observe the vastly larger and fascinating universe around us. (a) Side view. (b) View from above. (c) The Milky Way as seen from Earth. (credits: (a) NASA, (b) Nick Risinger, (c) Andy)

Distances are great even within our galaxy and are measured in light years (the distance traveled by light in one year). The average distance between galaxies is on the order of a million light years, but it varies greatly with galaxies forming clusters such as shown in [link]. The Magellanic Clouds, for example, are small galaxies close to our own, some 160,000 light years from Earth. The Andromeda galaxy is a large spiral galaxy like ours and lies 2 million light years away. It is just visible to the naked eye as an extended glow in the Andromeda constellation. Andromeda is the closest large galaxy in our local group, and we can see some individual stars in it with our larger telescopes. The most distant known galaxy is 14 billion light years from Earth—a truly incredible distance. (See [link].)


(a)
UDFj-39546284
Hubble Ultra Deep Field 2009-2010
Hubble Space Telescope • WFC3/IR
NASA, ESA, G. Illingworth and R. Bouwens (University of California, Santa Cruz), and the HUDF09 Team
STScI-PRC11-05
(b)

(a) Andromeda is the closest large galaxy, at 2 million light years distance, and is very similar to our Milky Way. The blue regions harbor young and emerging stars, while dark streaks are vast clouds of gas and dust. A smaller satellite galaxy is clearly visible. (b) The box indicates what may be the most distant known galaxy, estimated to be 13 billion light years from us. It exists in a much older part of the universe. (credit: NASA, ESA, G.

Illingworth (University of
California, Santa Cruz),
R. Bouwens (University
of California, Santa Cruz
and Leiden University),
and the HUDF09 Team)

Consider the fact that the light we receive from these vast distances has been on its way to us for a long time. In fact, the time in years is the same as the distance in light years. For example, the Andromeda galaxy is 2 million light years away, so that the light now reaching us left it 2 million years ago. If we could be there now, Andromeda would be different. Similarly, light from the most distant galaxy left it 14 billion years ago. We have an incredible view of the past when looking great distances. We can try to see if the universe was different then—if distant galaxies are more tightly packed or have younger-looking stars, for example, than closer galaxies, in which case there has been an evolution in time. But the problem is that the uncertainties in our data are great. Cosmology is almost typified by these large uncertainties, so that we must be especially cautious in drawing conclusions. One consequence is that there are more questions than answers, and so there are many competing theories. Another consequence is that any hard data produce a major result. Discoveries of some importance are being made on a regular basis, the hallmark of a field in its golden age.

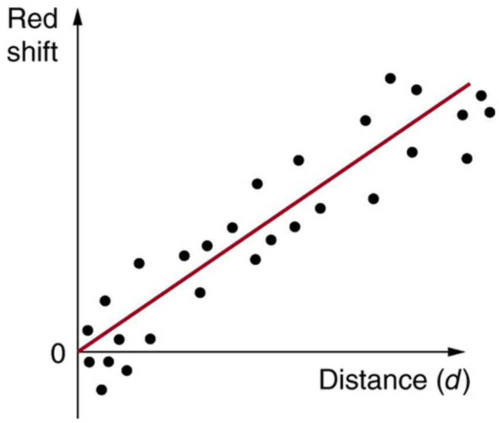
Perhaps the most important characteristic of the universe is that all galaxies except those in our local cluster seem to be moving away from us at speeds proportional to their distance from our galaxy. It looks as if a gigantic explosion, universally called the **Big Bang**, threw matter out some billions of years ago. This amazing conclusion is based on the pioneering work of Edwin Hubble (1889–1953), the American astronomer. In the 1920s, Hubble first demonstrated conclusively that other galaxies, many previously called nebulae or clouds of stars, were outside our own. He then found that all but the closest galaxies have a red shift in their hydrogen spectra that is proportional to their distance. The explanation is that there is a **cosmological red shift** due to the expansion of space itself. The photon wavelength is stretched in transit from the source to the observer. Double the distance, and the red shift is doubled. While this cosmological red shift is often called a Doppler shift, it is not—space itself is expanding. There is no center of expansion in the universe. All observers see themselves as stationary; the other objects in space appear to be moving away from them. Hubble was directly responsible for discovering that the universe was much larger than had previously been imagined and that it had this amazing characteristic of rapid expansion.

Universal expansion on the scale of galactic clusters (that is, galaxies at smaller distances are not uniformly receding from one another) is an integral part of modern cosmology. For galaxies farther away than about 50 Mly (50 million light years), the expansion is uniform with variations due to local motions of galaxies within clusters. A representative recession velocity v can be obtained from the simple formula
Equation:

$$\mathrm{v} = H_0 d,$$

where d is the distance to the galaxy and H_0 is the **Hubble constant**. The Hubble constant is a central concept in cosmology. Its value is determined by taking the slope of a graph of velocity versus distance, obtained from red shift measurements, such as shown in [link]. We shall use an approximate value of $H_0 = 20\ \text{km/s} \cdot\ \text{Mly}$. Thus, $v = H_0 d$ is an average behavior for all but the closest galaxies. For example, a galaxy 100 Mly away (as determined by its size and brightness) typically moves away from us at a speed of $v = (20\ \text{km/s} \cdot\ \text{Mly})(100\ \text{Mly}) = 2000\ \text{km/s}$. There can be variations in this speed due to so-called local motions or interactions with neighboring galaxies. Conversely, if a galaxy is found to be moving away from us at speed of 100,000 km/s based on its red shift, it is at a distance

$d = v/H_0 = (10{,}000\ \text{km/s})/(20\ \text{km/s} \cdot\ \text{Mly}) = 5000\ \text{Mly} = 5\ \text{Gly}$ or 5×10^9 ly. This last calculation is approximate, because it assumes the expansion rate was the same 5 billion years ago as now. A similar calculation in Hubble’s measurement changed the notion that the universe is in a steady state.

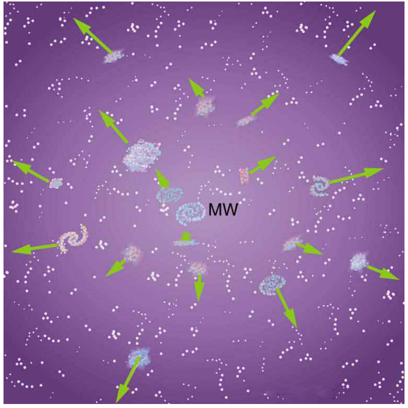


This graph of red shift versus distance for galaxies shows a linear relationship, with larger red shifts at greater distances, implying an expanding universe. The slope gives an approximate value for the expansion rate. (credit: John Cub).

One of the most intriguing developments recently has been the discovery that the expansion of the universe may be *faster now* than in the past, rather than slowing due to gravity as expected. Various groups have been looking, in particular, at supernovas in moderately distant galaxies (less than 1 Gly) to get improved distance measurements. Those distances are larger than expected for the observed galactic red shifts, implying the expansion was slower when that light was emitted. This has cosmological consequences that are discussed in Dark Matter and Closure. The first results, published in 1999, are only the beginning of emerging data, with astronomy now entering a data-rich era.

[link] shows how the recession of galaxies looks like the remnants of a gigantic explosion, the famous Big Bang. Extrapolating backward in time,

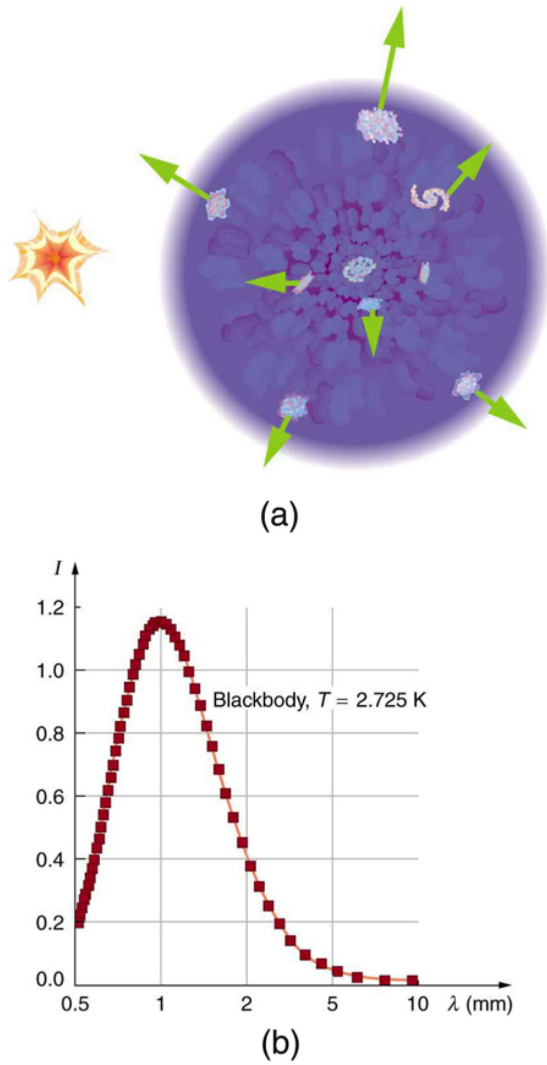
the Big Bang would have occurred between 13 and 15 billion years ago when all matter would have been at a point. Questions instantly arise. What caused the explosion? What happened before the Big Bang? Was there a before, or did time start then? Will the universe expand forever, or will gravity reverse it into a Big Crunch? And is there other evidence of the Big Bang besides the well-documented red shifts?



Galaxies are flying apart from one another, with the more distant moving faster as if a primordial explosion expelled the matter from which they formed. The most distant known galaxies move nearly at the speed of light relative to us.

The Russian-born American physicist George Gamow (1904–1968) was among the first to note that, if there was a Big Bang, the remnants of the primordial fireball should still be evident and should be blackbody radiation. Since the radiation from this fireball has been traveling to us since shortly after the Big Bang, its wavelengths should be greatly stretched. It will look as if the fireball has cooled in the billions of years since the Big Bang. Gamow and collaborators predicted in the late 1940s that there should be blackbody radiation from the explosion filling space with a characteristic temperature of about 7 K. Such blackbody radiation would have its peak intensity in the microwave part of the spectrum. (See [link].) In 1964, Arno Penzias and Robert Wilson, two American scientists working with Bell Telephone Laboratories on a low-noise radio antenna, detected the radiation and eventually recognized it for what it is.

[link](b) shows the spectrum of this microwave radiation that permeates space and is of cosmic origin. It is the most perfect blackbody spectrum known, and the temperature of the fireball remnant is determined from it to be $2.725 \pm 0.\ 002\ \text{K}$. The detection of what is now called the **cosmic microwave background** (CMBR) was so important (generally considered as important as Hubble’s detection that the galactic red shift is proportional to distance) that virtually every scientist has accepted the expansion of the universe as fact. Penzias and Wilson shared the 1978 Nobel Prize in Physics for their discovery.



(a) The Big Bang is used to explain the present observed expansion of the universe. It was an incredibly energetic explosion some 10 to 20 billion years ago. After expanding and cooling, galaxies form inside the now-cold remnants of the primordial fireball. (b) The spectrum of cosmic microwave radiation is the most perfect blackbody spectrum ever detected. It is characteristic of a temperature of 2.725 K, the expansion-cooled temperature of the Big Bang’s remnant. This radiation can be measured coming from any direction in space not obscured by some other source. It is compelling evidence of the creation of the universe in a gigantic explosion, already indicated by galactic red shifts.

Note:
Making Connections: Cosmology and Particle Physics
There are many connections of cosmology—by definition involving physics on the largest scale—with particle physics—by definition physics on the smallest scale. Among these are the dominance of matter over antimatter, the nearly perfect uniformity of the cosmic microwave background, and the mere existence of galaxies.

Matter versus antimatter

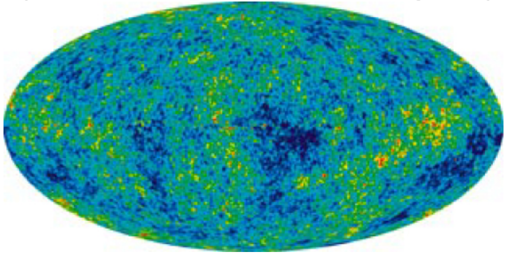
We know from direct observation that antimatter is rare. The Earth and the solar system are nearly pure matter. Space probes and cosmic rays give direct evidence—the landing of the Viking probes on Mars would have been spectacular explosions of mutual annihilation energy if Mars were antimatter. We also know that most of the universe is dominated by matter. This is proven by the lack of annihilation radiation coming to us from space, particularly the relative absence of 0.511-MeV γ rays created by the mutual annihilation of electrons and positrons. It seemed possible that there could be entire solar systems or galaxies made of antimatter in perfect symmetry with our matter-dominated systems. But the interactions between stars and galaxies would sometimes bring matter and antimatter together in large amounts. The annihilation radiation they would produce is simply not observed. Antimatter in nature is created in particle collisions and in β^+ decays, but only in small amounts that quickly annihilate, leaving almost pure matter surviving.

Particle physics seems symmetric in matter and antimatter. Why isn't the cosmos? The answer is that particle physics is not quite perfectly symmetric in this regard. The decay of one of the neutral K-mesons, for example, preferentially creates more matter than antimatter. This is caused by a fundamental small asymmetry in the basic forces. This small asymmetry produced slightly more matter than antimatter in the early universe. If there was only one part in 10^9 more matter (a small asymmetry), the rest would annihilate pair for pair, leaving nearly pure matter to form the stars and galaxies we see today. So the vast number of stars we observe may be only a tiny remnant of the original matter created in the Big Bang. Here at last we see a very real and important asymmetry in nature. Rather than be disturbed by an asymmetry, most physicists are impressed by how small it is. Furthermore, if the universe were completely symmetric, the mutual annihilation would be more complete, leaving far less matter to form us and the universe we know.

How can something so old have so few wrinkles?

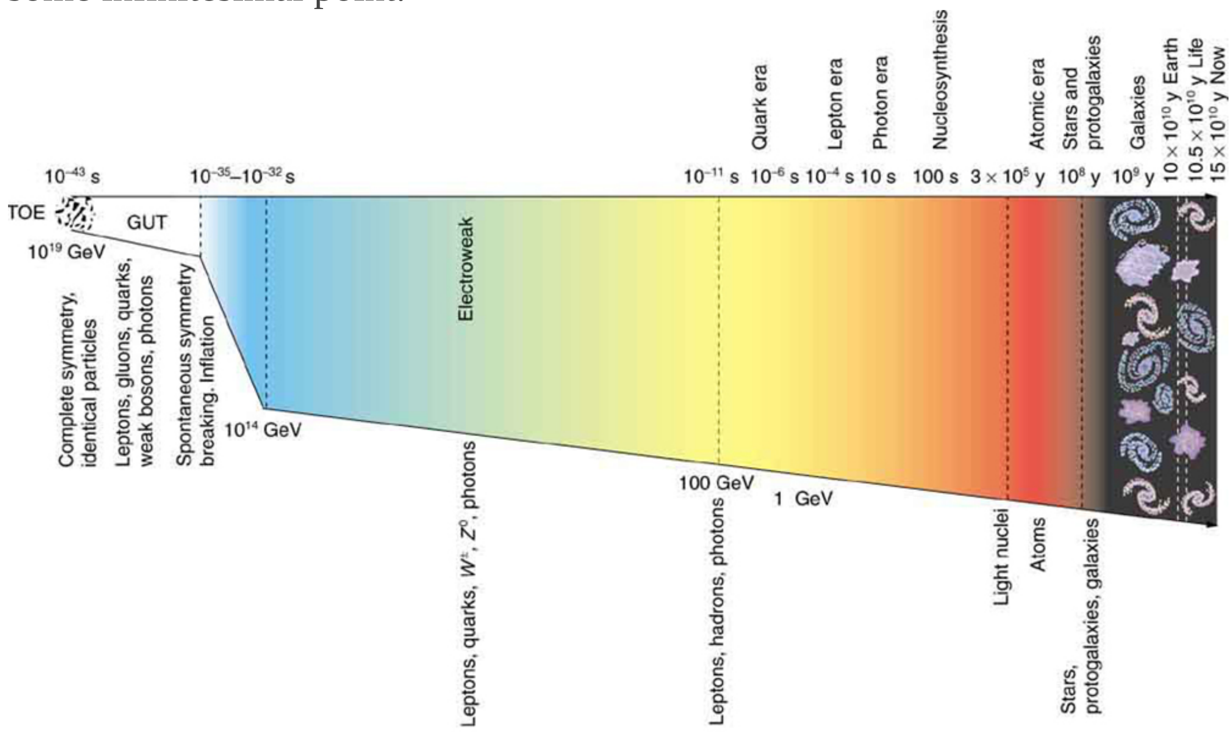
A troubling aspect of cosmic microwave background radiation (CMBR) was soon recognized. True, the CMBR verified the Big Bang, had the correct temperature, and had a blackbody spectrum as expected. But the CMBR was *too* smooth—it looked identical in every direction. Galaxies and other similar entities could not be formed without the existence of fluctuations in the primordial stages of the universe and so there should be hot and cool spots in the CMBR, nicknamed wrinkles, corresponding to dense and sparse regions of gas caused by turbulence or early fluctuations. Over time, dense regions would contract under gravity and form stars and galaxies. Why aren't the fluctuations there? (This is a good example of an answer producing more questions.) Furthermore, galaxies are observed very far from us, so that they formed very long ago. The problem was to explain how galaxies could form so early and so quickly after the Big Bang if its remnant fingerprint is perfectly smooth. The answer is that if you look very closely, the CMBR is not perfectly smooth, only extremely smooth.

A satellite called the Cosmic Background Explorer (COBE) carried an instrument that made very sensitive and accurate measurements of the CMBR. In April of 1992, there was extraordinary publicity of COBE’s first results—there were small fluctuations in the CMBR. Further measurements were carried out by experiments including NASA’s Wilkinson Microwave Anisotropy Probe (WMAP), which launched in 2001. Data from WMAP provided a much more detailed picture of the CMBR fluctuations. (See [link].) These amount to temperature fluctuations of only $200\ \mu\text{k}$ out of 2.7 K, better than one part in 1000. The WMAP experiment will be followed up by the European Space Agency’s Planck Surveyor, which launched in 2009.



This map of the sky uses color to show fluctuations, or wrinkles, in the cosmic microwave background observed with the WMAP spacecraft. The Milky Way has been removed for clarity. Red represents higher temperature and higher density, while blue is lower temperature and density. The fluctuations are small, less than one part in 1000, but these are still thought to be the cause of the eventual formation of galaxies. (credit: NASA/WMAP Science Team)

Let us now examine the various stages of the overall evolution of the universe from the Big Bang to the present, illustrated in [link]. Note that scientific notation is used to encompass the many orders of magnitude in time, energy, temperature, and size of the universe. Going back in time, the two lines approach but do not cross (there is no zero on an exponential scale). Rather, they extend indefinitely in ever-smaller time intervals to some infinitesimal point.


Quark era
Lepton era
Photon era
Nucleosynthesis
Atomic era
Stars and protogalaxies
Galaxies
10×10^{10} y Earth
10.5×10^{10} y Life
15×10^{10} y Now
10^{-43} s
10^{-35}–10^{-32} s
10^{-11} s
10^{-6} s
10^{-4} s
10 s
100 s
3×10^5 y
10^8 y
10^9 y
TOE
GUT
10^{19} GeV
Electroweak
10^{14} GeV
100 GeV
1 GeV
Complete symmetry, identical particles
Leptons, gluons, quarks, weak bosons, photons
Spontaneous symmetry breaking. Inflation
Leptons, quarks, $W^{\pm}$, Z^0, photons
Leptons, hadrons, photons
Light nuclei
Atoms
Stars, protogalaxies, galaxies

The evolution of the universe from the Big Bang onward is intimately tied to the laws of physics, especially those of particle physics at the earliest stages. The universe is relativistic throughout its history. Theories of the unification of forces at high energies may be verified by their shaping of the universe and its evolution.

Going back in time is equivalent to what would happen if expansion stopped and gravity pulled all the galaxies together, compressing and heating all matter. At a time long ago, the temperature and density were too high for stars and galaxies to exist. Before then, there was a time when the temperature was too great for atoms to exist. And farther back yet, there was a time when the temperature and density were so great that nuclei could not exist. Even farther back in time, the temperature was so high that average kinetic energy was great enough to create short-lived particles, and the density was high enough to make this likely. When we extrapolate back to the point of $W^{\pm}$ and Z^0 production (thermal energies reaching 1 TeV, or a temperature of about 10^{15} K), we reach the limits of what we know directly about particle physics. This is at a time about 10^{-12} s after the Big Bang. While 10^{-12} s may seem to be negligibly close to the instant of creation, it is not. There are important stages before this time that are tied to the unification of forces. At those stages, the universe was at extremely high energies and average particle separations were smaller than we can achieve with accelerators. What happened in the early stages before 10^{-12} s is crucial to all later stages and is possibly discerned by observing present conditions in the universe. One of these is the smoothness of the CMBR.

Names are given to early stages representing key conditions. The stage before 10^{-11} s back to 10^{-34} s is called the **electroweak epoch**, because the electromagnetic and weak forces become identical for energies above about 100 GeV. As discussed earlier, theorists expect that the strong force becomes identical to and thus unified with the electroweak force at energies of about 10^{14} GeV. The average particle energy would be this great at 10^{-34} s after the Big Bang, if there are no surprises in the unknown physics at energies above about 1 TeV. At the immense energy of 10^{14} GeV (corresponding to a temperature of about 10^{26} K), the $W^{\pm}$ and Z^0 carrier particles would be transformed into massless gauge bosons to accomplish the unification. Before 10^{-34} s back to about 10^{-43} s, we have Grand Unification in the **GUT epoch**, in which all forces except gravity are identical. At 10^{-43} s, the average energy reaches the immense 10^{19} GeV needed to unify gravity with the other forces in TOE, the Theory of Everything. Before that time is the **TOE epoch**, but we have almost no idea as to the nature of the universe then, since we have no workable theory of quantum gravity. We call the hypothetical unified force **superforce**.

Now let us imagine starting at TOE and moving forward in time to see what type of universe is created from various events along the way. As temperatures and average energies decrease with expansion, the universe reaches the stage where average particle separations are large enough to see differences between the strong and electroweak forces (at about 10^{-35} s). After this time, the forces become distinct in almost all interactions—they are no longer unified or symmetric. This transition from GUT to electroweak is an example of **spontaneous symmetry breaking**, in which conditions spontaneously evolved to a point where the forces were no longer unified, breaking that symmetry. This is analogous to a phase transition in the universe, and a clever proposal by American physicist Alan Guth in the early 1980s ties it to the smoothness of the CMBR. Guth proposed that spontaneous symmetry breaking (like a phase transition during cooling of normal matter) released an immense amount of energy that caused the universe to expand extremely rapidly for the brief time from 10^{-35} s to about 10^{-32} s. This expansion may have been by an incredible factor of 10^{50} or more in the size of the universe and is thus called the **inflationary scenario**. One result of this inflation is that it would stretch the wrinkles in the universe nearly flat, leaving an extremely smooth CMBR. While speculative, there is as yet no other plausible explanation for the smoothness of the CMBR. Unless the CMBR is not really cosmic but local in origin, the distances between regions of similar temperatures are too great for any coordination to have caused them, since any coordination mechanism must travel at the speed of light. Again, particle physics and cosmology are intimately entwined. There is little hope that we may be able to test the inflationary scenario directly, since it occurs at energies near 10^{14} GeV, vastly greater than the limits of modern accelerators. But the idea is so attractive that it is incorporated into most cosmological theories.

Characteristics of the present universe may help us determine the validity of this intriguing idea. Additionally, the recent indications that the universe’s expansion rate may be *increasing* (see Dark Matter and Closure) could even imply that we are *in* another inflationary epoch.

It is important to note that, if conditions such as those found in the early universe could be created in the laboratory, we would see the unification of forces directly today. The forces have not changed in time, but the average energy and separation of particles in the universe have. As discussed in The Four Basic Forces, the four basic forces in nature are distinct under most circumstances found today. The early universe and its remnants provide evidence from times when they were unified under most circumstances.

Section Summary

- Cosmology is the study of the character and evolution of the universe.
- The two most important features of the universe are the cosmological red shifts of its galaxies being proportional to distance and its cosmic microwave background (CMBR). Both support the notion that there was a gigantic explosion, known as the Big Bang that created the universe.
- Galaxies farther away than our local group have, on an average, a recessional velocity given by
  Equation:

  $$v = H_0 d,$$

  where d is the distance to the galaxy and H_0 is the Hubble constant, taken to have the average value $H_0 = 20\ \text{km/s} \cdot \text{Mly}$.
- Explanations of the large-scale characteristics of the universe are intimately tied to particle physics.
- The dominance of matter over antimatter and the smoothness of the CMBR are two characteristics that are tied to particle physics.
- The epochs of the universe are known back to very shortly after the Big Bang, based on known laws of physics.
- The earliest epochs are tied to the unification of forces, with the electroweak epoch being partially understood, the GUT epoch being speculative, and the TOE epoch being highly speculative since it involves an unknown single superforce.
- The transition from GUT to electroweak is called spontaneous symmetry breaking. It released energy that caused the inflationary

scenario, which in turn explains the smoothness of the CMBR.

Conceptual Questions

Exercise:

Problem:

Explain why it only *appears* that we are at the center of expansion of the universe and why an observer in another galaxy would see the same relative motion of all but the closest galaxies away from her.

Exercise:

Problem:

If there is no observable edge to the universe, can we determine where its center of expansion is? Explain.

Exercise:

Problem: If the universe is infinite, does it have a center? Discuss.

Exercise:

Problem:

Another known cause of red shift in light is the source being in a high gravitational field. Discuss how this can be eliminated as the source of galactic red shifts, given that the shifts are proportional to distance and not to the size of the galaxy.

Exercise:

Problem:

If some unknown cause of red shift—such as light becoming “tired” from traveling long distances through empty space—is discovered, what effect would there be on cosmology?

Exercise:

Problem:

Olbers’s paradox poses an interesting question: If the universe is infinite, then any line of sight should eventually fall on a star’s surface. Why then is the sky dark at night? Discuss the commonly accepted evolution of the universe as a solution to this paradox.

Exercise:

Problem:

If the cosmic microwave background radiation (CMBR) is the remnant of the Big Bang’s fireball, we expect to see hot and cold regions in it. What are two causes of these wrinkles in the CMBR? Are the observed temperature variations greater or less than originally expected?

Exercise:

Problem:

The decay of one type of K-meson is cited as evidence that nature favors matter over antimatter. Since mesons are composed of a quark and an antiquark, is it surprising that they would preferentially decay to one type over another? Is this an asymmetry in nature? Is the predominance of matter over antimatter an asymmetry?

Exercise:

Problem:

Distances to local galaxies are determined by measuring the brightness of stars, called Cepheid variables, that can be observed individually and that have absolute brightnesses at a standard distance that are well known. Explain how the measured brightness would vary with distance as compared with the absolute brightness.

Exercise:

Problem:

Distances to very remote galaxies are estimated based on their apparent type, which indicate the number of stars in the galaxy, and their measured brightness. Explain how the measured brightness would vary with distance. Would there be any correction necessary to compensate for the red shift of the galaxy (all distant galaxies have significant red shifts)? Discuss possible causes of uncertainties in these measurements.

Exercise:

Problem:

If the smallest meaningful time interval is greater than zero, will the lines in [link] ever meet?

Problems & Exercises

Exercise:

Problem:

Find the approximate mass of the luminous matter in the Milky Way galaxy, given it has approximately 10^{11} stars of average mass 1.5 times that of our Sun.

Solution:

3×10^{41} kg

Exercise:

Problem:

Find the approximate mass of the dark and luminous matter in the Milky Way galaxy. Assume the luminous matter is due to approximately 10^{11} stars of average mass 1.5 times that of our Sun, and take the dark matter to be 10 times as massive as the luminous matter.

Exercise:

Problem:

(a) Estimate the mass of the luminous matter in the known universe, given there are 10^{11} galaxies, each containing 10^{11} stars of average mass 1.5 times that of our Sun. (b) How many protons (the most abundant nuclide) are there in this mass? (c) Estimate the total number of particles in the observable universe by multiplying the answer to (b) by two, since there is an electron for each proton, and then by 10^{9}, since there are far more particles (such as photons and neutrinos) in space than in luminous matter.

Solution:

(a) 3×10^{52} kg

(b) 2×10^{79}

(c) 4×10^{88}

Exercise:

Problem:

If a galaxy is 500 Mly away from us, how fast do we expect it to be moving and in what direction?

Exercise:

Problem:

On average, how far away are galaxies that are moving away from us at 2.0% of the speed of light?

Solution:

0.30 Gly

Exercise:

Problem:

Our solar system orbits the center of the Milky Way galaxy. Assuming a circular orbit 30,000 ly in radius and an orbital speed of 250 km/s, how many years does it take for one revolution? Note that this is approximate, assuming constant speed and circular orbit, but it is representative of the time for our system and local stars to make one revolution around the galaxy.

Exercise:

Problem:

(a) What is the approximate speed relative to us of a galaxy near the edge of the known universe, some 10 Gly away? (b) What fraction of the speed of light is this? Note that we have observed galaxies moving away from us at greater than $0.9c$.

Solution:

(a) $2.0 \times 10^5 \text{ km/s}$

(b) $0.67c$

Exercise:

Problem:

(a) Calculate the approximate age of the universe from the average value of the Hubble constant, $H_0 = 20\ \text{km/s}$ ·Mly. To do this, calculate the time it would take to travel 1 Mly at a constant expansion rate of 20 km/s. (b) If deceleration is taken into account, would the actual age of the universe be greater or less than that found here? Explain.

Exercise:

Problem:

Assuming a circular orbit for the Sun about the center of the Milky Way galaxy, calculate its orbital speed using the following information: The mass of the galaxy is equivalent to a single mass 1.5×10^{11} times that of the Sun (or 3×10^{41} kg), located 30,000 ly away.

Solution:

$2.7 \times 10^5\ \text{m/s}$

Exercise:

Problem:

(a) What is the approximate force of gravity on a 70-kg person due to the Andromeda galaxy, assuming its total mass is 10^{13} that of our Sun and acts like a single mass 2 Mly away? (b) What is the ratio of this force to the person’s weight? Note that Andromeda is the closest large galaxy.

Exercise:

Problem:

Andromeda galaxy is the closest large galaxy and is visible to the naked eye. Estimate its brightness relative to the Sun, assuming it has luminosity 10^{12} times that of the Sun and lies 2 Mly away.

Solution:

6×10^{-11} (an overestimate, since some of the light from Andromeda is blocked by gas and dust within that galaxy)

Exercise:

Problem:

(a) A particle and its antiparticle are at rest relative to an observer and annihilate (completely destroying both masses), creating two γ rays of equal energy. What is the characteristic γ-ray energy you would look for if searching for evidence of proton-antiproton annihilation? (The fact that such radiation is rarely observed is evidence that there is very little antimatter in the universe.) (b) How does this compare with the 0.511-MeV energy associated with electron-positron annihilation?

Exercise:

Problem:

The average particle energy needed to observe unification of forces is estimated to be $10^{19}\ \text{GeV}$. (a) What is the rest mass in kilograms of a particle that has a rest mass of $10^{19}\ \text{GeV}/c^2$? (b) How many times the mass of a hydrogen atom is this?

Solution:

(a) $2 \times 10^{-8}\ \text{kg}$

(b) 1×10^{19}

Exercise:

Problem:

The peak intensity of the CMBR occurs at a wavelength of 1.1 mm. (a) What is the energy in eV of a 1.1-mm photon? (b) There are approximately 10^9 photons for each massive particle in deep space. Calculate the energy of 10^9 such photons. (c) If the average massive particle in space has a mass half that of a proton, what energy would be created by converting its mass to energy? (d) Does this imply that space is “matter dominated”? Explain briefly.

Exercise:

Problem:

(a) What Hubble constant corresponds to an approximate age of the universe of 10^{10} y? To get an approximate value, assume the expansion rate is constant and calculate the speed at which two galaxies must move apart to be separated by 1 Mly (present average galactic separation) in a time of 10^{10} y. (b) Similarly, what Hubble constant corresponds to a universe approximately 2×10^{10}-y old?

Solution:

(a) $30\text{ km/s}\cdot\text{Mly}$

(b) $15\text{ km/s}\cdot\text{Mly}$

Exercise:

Problem:

Show that the velocity of a star orbiting its galaxy in a circular orbit is inversely proportional to the square root of its orbital radius, assuming the mass of the stars inside its orbit acts like a single mass at the center of the galaxy. You may use an equation from a previous chapter to support your conclusion, but you must justify its use and define all terms used.

Exercise:

Problem:

The core of a star collapses during a supernova, forming a neutron star. Angular momentum of the core is conserved, and so the neutron star spins rapidly. If the initial core radius is $5.0\times 10^5\ \text{km}$ and it collapses to 10.0 km, find the neutron star's angular velocity in revolutions per second, given the core's angular velocity was originally 1 revolution per 30.0 days.

Solution:

960 rev/s

Exercise:

Problem:

Using data from the previous problem, find the increase in rotational kinetic energy, given the core's mass is 1.3 times that of our Sun. Where does this increase in kinetic energy come from?

Exercise:

Problem:

Distances to the nearest stars (up to 500 ly away) can be measured by a technique called parallax, as shown in [link]. What are the angles θ_1 and θ_2 relative to the plane of the Earth's orbit for a star 4.0 ly directly above the Sun?

Solution:

89.999773º (many digits are used to show the difference between 90º)

Exercise:

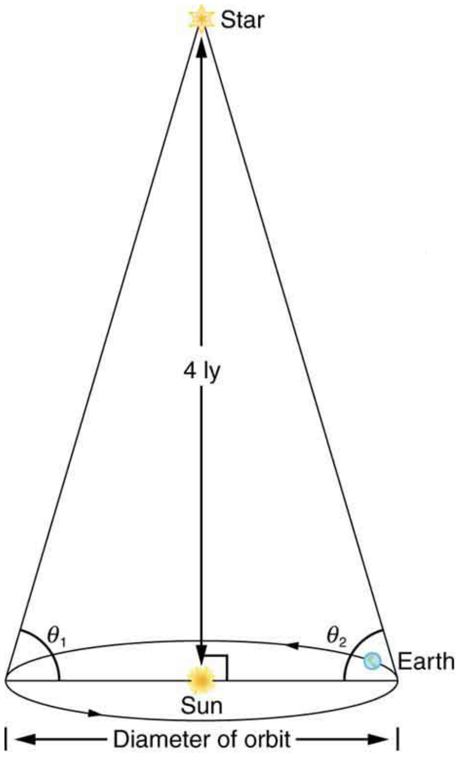
Problem:

(a) Use the Heisenberg uncertainty principle to calculate the uncertainty in energy for a corresponding time interval of 10^{-43} s. (b) Compare this energy with the 10^{19} GeV unification-of-forces energy and discuss why they are similar.

Exercise:

Problem: Construct Your Own Problem

Consider a star moving in a circular orbit at the edge of a galaxy. Construct a problem in which you calculate the mass of that galaxy in kg and in multiples of the solar mass based on the velocity of the star and its distance from the center of the galaxy.



Distances to nearby stars are measured using triangulation, also called the parallax method. The angle of line of sight to the star is measured at intervals six months apart, and the distance is calculated by using the known diameter of the Earth’s orbit. This can be done for stars up to about 500 ly away.

Glossary

Big Bang
a gigantic explosion that threw out matter a few billion years ago

cosmic microwave background
the spectrum of microwave radiation of cosmic origin

cosmological red shift
the photon wavelength is stretched in transit from the source to the observer because of the expansion of space itself

cosmology
the study of the character and evolution of the universe

electroweak epoch
the stage before 10^{-11} back to 10^{-34} after the Big Bang

GUT epoch
the time period from 10^{-43} to 10^{-34} after the Big Bang, when Grand Unification Theory, in which all forces except gravity are identical, governed the universe

Hubble constant

a central concept in cosmology whose value is determined by taking the slope of a graph of velocity versus distance, obtained from red shift measurements

inflationary scenario

the rapid expansion of the universe by an incredible factor of 10^{-50} for the brief time from 10^{-35} to about 10^{-32}s

spontaneous symmetry breaking

the transition from GUT to electroweak where the forces were no longer unified

superforce

hypothetical unified force in TOE epoch

TOE epoch

before 10^{-43} after the Big Bang

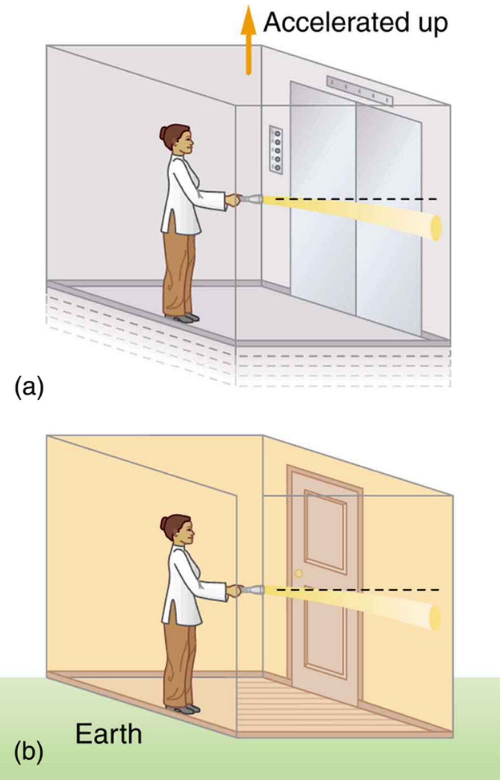
General Relativity and Quantum Gravity

- Explain the effect of gravity on light.
- Discuss black hole.
- Explain quantum gravity.

When we talk of black holes or the unification of forces, we are actually discussing aspects of general relativity and quantum gravity. We know from Special Relativity that relativity is the study of how different observers measure the same event, particularly if they move relative to one another. Einstein’s theory of **general relativity** describes all types of relative motion including accelerated motion and the effects of gravity. General relativity encompasses special relativity and classical relativity in situations where acceleration is zero and relative velocity is small compared with the speed of light. Many aspects of general relativity have been verified experimentally, some of which are better than science fiction in that they are bizarre but true. **Quantum gravity** is the theory that deals with particle exchange of gravitons as the mechanism for the force, and with extreme conditions where quantum mechanics and general relativity must both be used. A good theory of quantum gravity does not yet exist, but one will be needed to understand how all four forces may be unified. If we are successful, the theory of quantum gravity will encompass all others, from classical physics to relativity to quantum mechanics—truly a Theory of Everything (TOE).

General Relativity

Einstein first considered the case of no observer acceleration when he developed the revolutionary special theory of relativity, publishing his first work on it in 1905. By 1916, he had laid the foundation of general relativity, again almost on his own. Much of what Einstein did to develop his ideas was to mentally analyze certain carefully and clearly defined situations—doing this is to perform a **thought experiment**. [link] illustrates a thought experiment like the ones that convinced Einstein that light must fall in a gravitational field. Think about what a person feels in an elevator that is accelerated upward. It is identical to being in a stationary elevator in a gravitational field. The feet of a person are pressed against the floor, and objects released from hand fall with identical accelerations. In fact, it is not possible, without looking outside, to know what is happening—acceleration upward or gravity. This led Einstein to correctly postulate that acceleration and gravity will produce identical effects in all situations. So, if acceleration affects light, then gravity will, too. [link] shows the effect of acceleration on a beam of light shone horizontally at one wall. Since the accelerated elevator moves up during the time light travels across the elevator, the beam of light strikes low, seeming to the person to bend down. (Normally a tiny effect, since the speed of light is so great.) The same effect must occur due to gravity, Einstein reasoned, since there is no way to tell the effects of gravity acting downward from acceleration of the elevator upward. Thus gravity affects the path of light, even though we think of gravity as acting between masses and photons are massless.



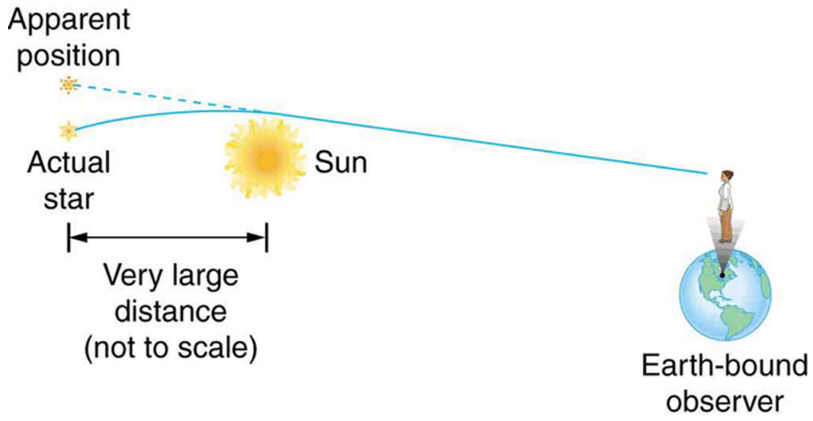
(a) A beam of light emerges from a flashlight in an upward-accelerating elevator. Since the elevator moves up during the time the light takes to reach the wall, the beam strikes lower than it would if the elevator were not accelerated. (b) Gravity has the same effect on light, since it is not possible to tell whether the elevator is accelerating upward or acted upon by gravity.

Einstein’s theory of general relativity got its first verification in 1919 when starlight passing near the Sun was observed during a solar eclipse. (See [link].) During an eclipse, the sky is darkened and we can briefly see stars. Those in a line of sight nearest the Sun should have a shift in their apparent positions. Not only was this shift observed, but it agreed with Einstein’s predictions well within experimental uncertainties. This discovery created a scientific and public sensation. Einstein was now a folk hero as well as a very great scientist. The bending of light by matter is equivalent to a bending of space itself, with light following the curve. This is another radical change in our concept of space and time. It is also another connection that any particle with mass or energy (massless photons) is affected by gravity.

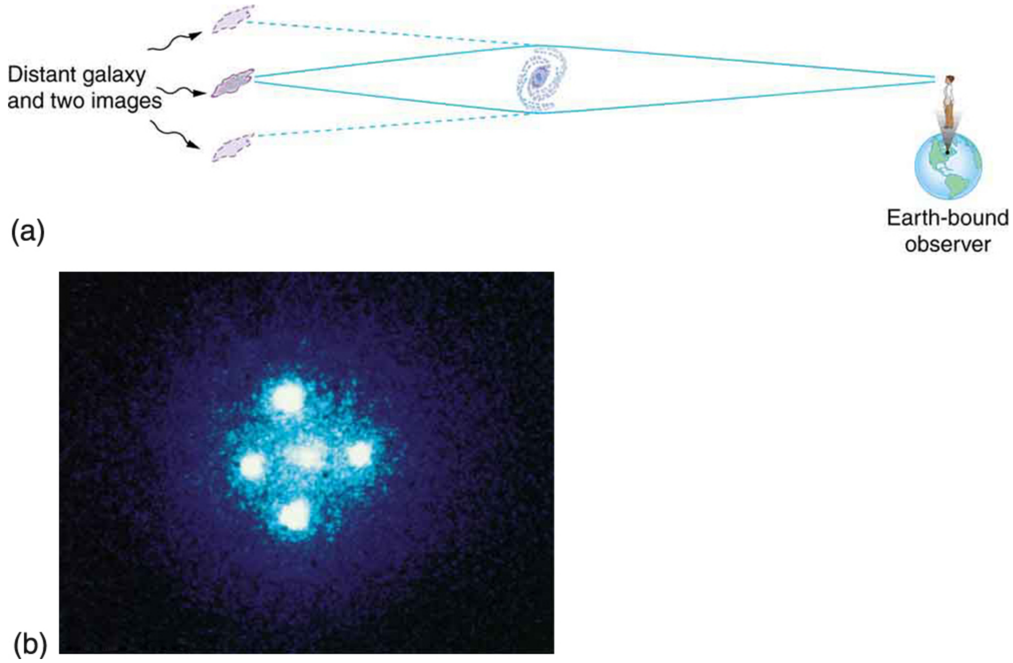
There are several current forefront efforts related to general relativity. One is the observation and analysis of gravitational lensing of light. Another is analysis of the definitive proof of the existence of black holes. Direct observation of gravitational waves or moving wrinkles in space is being searched for. Theoretical efforts are also being aimed at the possibility of time travel and wormholes into other parts of space due to black holes.

Gravitational lensing

As you can see in [link], light is bent toward a mass, producing an effect much like a converging lens (large masses are needed to produce observable effects). On a galactic scale, the light from a distant galaxy could be “lensed” into several images when passing close by another galaxy on its way to Earth. Einstein predicted this effect, but he considered it unlikely that we would ever observe it. A number of cases of this effect have now been observed; one is shown in [link]. This effect is a much larger scale verification of general relativity. But such gravitational lensing is also useful in verifying that the red shift is proportional to distance. The red shift of the intervening galaxy is always less than that of the one being lensed, and each image of the lensed galaxy has the same red shift. This verification supplies more evidence that red shift is proportional to distance. Confidence that the multiple images are not different objects is bolstered by the observations that if one image varies in brightness over time, the others also vary in the same manner.



This schematic shows how light passing near a massive body like the Sun is curved toward it. The light that reaches the Earth then seems to be coming from different locations than the known positions of the originating stars. Not only was this effect observed, the amount of bending was precisely what Einstein predicted in his general theory of relativity.



(a) Light from a distant galaxy can travel different paths to the Earth because it is bent around an intermediary galaxy by gravity. This produces several images of the more distant galaxy. (b) The images around the central galaxy are produced by gravitational lensing. Each image has the same spectrum and a larger red shift than the intermediary. (credit: NASA, ESA, and STScI)

Black holes

Black holes are objects having such large gravitational fields that things can fall in, but nothing, not even light, can escape. Bodies, like the Earth or the Sun, have what is called an **escape velocity**. If an object moves straight up from the body, starting at the escape velocity, it will just be able to escape the gravity of the body. The greater the acceleration of gravity on the body, the greater is the escape velocity. As long ago as the late 1700s, it was proposed that if the escape velocity is greater than the speed of light, then light cannot escape. Simon Laplace (1749–1827), the French astronomer and mathematician, even incorporated this idea of a dark star into his writings. But the idea was dropped after Young's double slit experiment showed light to be a wave. For some time, light was thought not to have particle characteristics and, thus, could not be acted upon by gravity. The idea of a black hole was very quickly reincarnated in 1916 after Einstein's theory of general relativity was published. It is now thought that black holes can form in the supernova collapse of a massive star, forming an object perhaps 10 km across and having a mass greater than that of our Sun. It is interesting that several prominent physicists who worked on the concept, including Einstein, firmly believed that nature would find a way to prohibit such objects.

Black holes are difficult to observe directly, because they are small and no light comes directly from them. In fact, no light comes from inside the **event horizon**, which is defined to be at a distance from the object at which the escape velocity is exactly the speed of light. The radius of the event horizon is known as the **Schwarzschild radius** R_{S} and is given by

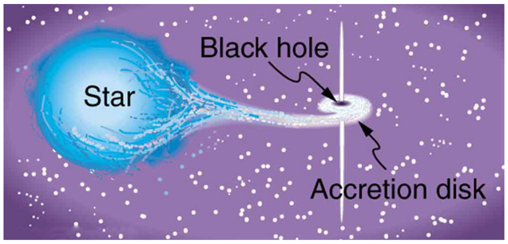
Equation:

$$R_{\text{S}} = \frac{2\text{GM}}{c^2},$$

where G is the universal gravitational constant, M is the mass of the body, and c is the speed of light. The event horizon is the edge of the black hole and R_{S} is its radius (that is, the size of a black hole is twice R_{S}). Since G is small and c^2 is large, you can see that black holes are extremely small, only a few kilometers for masses a little greater than the Sun's. The object itself is inside the event horizon.

Physics near a black hole is fascinating. Gravity increases so rapidly that, as you approach a black hole, the tidal effects tear matter apart, with matter closer to the hole being pulled in with much more force than that only slightly farther away. This can pull a companion star apart and heat inflowing gases to the point of producing X rays. (See [link].) We have observed X rays from certain binary star systems that are consistent with such a picture. This is not quite proof of black holes, because the X rays could also be caused by matter falling onto a neutron star. These objects were first discovered in 1967 by the British astrophysicists, Jocelyn Bell and Anthony Hewish. **Neutron stars** are literally a star composed of neutrons. They are formed by the collapse of a star's core in a supernova, during which electrons and protons are forced together to form neutrons (the reverse of neutron β decay). Neutron stars are slightly larger than a black hole of the same mass and will not collapse further because of resistance by the strong force. However, neutron stars cannot have a mass greater than about eight solar masses or they must collapse to a black hole. With recent improvements in our ability to resolve small details, such as with the orbiting Chandra X-ray Observatory, it has become possible to measure the masses of X-ray-emitting objects by observing the motion of companion stars and other matter in their vicinity. What has emerged is a plethora of X-ray-emitting objects too massive to be neutron stars. This evidence is considered conclusive and the existence of black holes is widely accepted. These black holes are concentrated near galactic centers.

We also have evidence that supermassive black holes may exist at the cores of many galaxies, including the Milky Way. Such a black hole might have a mass millions or even billions of times that of the Sun, and it would probably have formed when matter first coalesced into a galaxy billions of years ago. Supporting this is the fact that very distant galaxies are more likely to have abnormally energetic cores. Some of the moderately distant galaxies, and hence among the younger, are known as **quasars** and emit as much or more energy than a normal galaxy but from a region less than a light year across. Quasar energy outputs may vary in times less than a year, so that the energy-emitting region must be less than a light year across. The best explanation of quasars is that they are young galaxies with a supermassive black hole forming at their core, and that they become less energetic over billions of years. In closer superactive galaxies, we observe tremendous amounts of energy being emitted from very small regions of space, consistent with stars falling into a black hole at the rate of one or more a month. The Hubble Space Telescope (1994) observed an accretion disk in the galaxy M87 rotating rapidly around a region of extreme energy emission. (See [link].) A jet of material being ejected perpendicular to the plane of rotation gives further evidence of a supermassive black hole as the engine.



A black hole is shown pulling matter away from a companion star, forming a superheated accretion disk where X rays are emitted before the matter disappears forever into the hole. The in-fall energy also ejects some material, forming the two vertical spikes. (See also the photograph in Introduction to Frontiers of Physics.) There are several X-ray-emitting objects in space that are consistent with this picture and are likely to be black holes.

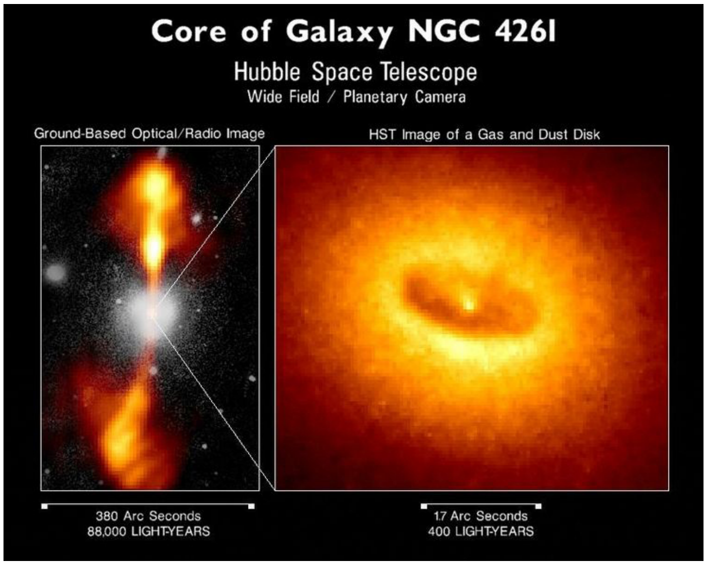
Gravitational waves

If a massive object distorts the space around it, like the foot of a water bug on the surface of a pond, then movement of the massive object should create waves in space like those on a pond. **Gravitational waves** are mass-created distortions in space that propagate at the speed of light and are predicted by general relativity. Since gravity is by far the weakest force, extreme conditions are needed to generate significant gravitational waves. Gravity near binary neutron star systems is so great that significant gravitational wave energy is radiated as the two neutron stars orbit one another. American astronomers, Joseph Taylor and Russell Hulse, measured changes in the orbit of such a binary neutron star system. They found its orbit to change precisely as predicted by general relativity, a strong indication of gravitational waves, and were awarded the 1993 Nobel Prize. But direct detection of gravitational waves on Earth would be conclusive. For many years, various attempts have been made to detect gravitational waves by observing vibrations induced in matter distorted by these waves. American physicist Joseph Weber pioneered this field in the 1960s, but no conclusive events have been observed. (No gravity wave detectors were in operation at the time of the 1987A supernova, unfortunately.) There are now several ambitious systems of gravitational wave detectors in use around the world. These include the LIGO (Laser Interferometer Gravitational Wave Observatory) system with two laser interferometer detectors, one in the state of Washington and another in Louisiana (See [link]) and the VIRGO (Variability of Irradiance and Gravitational Oscillations) facility in Italy with a single detector.

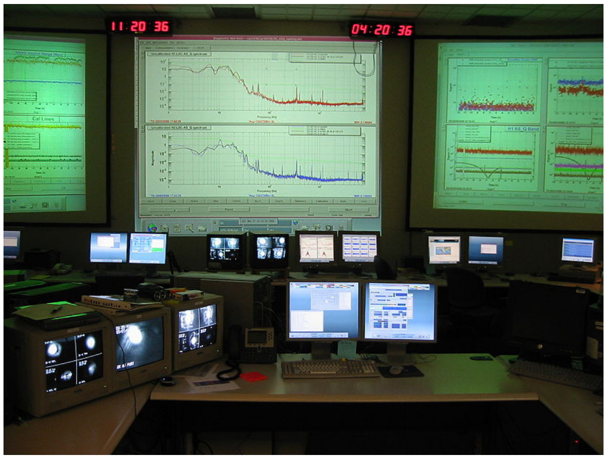
Quantum Gravity

Black holes radiate

Quantum gravity is important in those situations where gravity is so extremely strong that it has effects on the quantum scale, where the other forces are ordinarily much stronger. The early universe was such a place, but black holes are another. The first significant connection between gravity and quantum effects was made by the Russian physicist Yakov Zel’dovich in 1971, and other significant advances followed from the British physicist Stephen Hawking. (See [link].) These two showed that black holes could radiate away energy by quantum effects just outside the event horizon (nothing can escape from inside the event horizon). Black holes are, thus, expected to radiate energy and shrink to nothing, although extremely slowly for most black holes. The mechanism is the creation of a particle-antiparticle pair from energy in the extremely strong gravitational field near the event horizon. One member of the pair falls into the hole and the other escapes, conserving momentum. (See [link].) When a black hole loses energy and, hence, rest mass, its event horizon shrinks, creating an even greater gravitational field. This increases the rate of pair production so that the process grows exponentially until the black hole is nuclear in size. A final burst of particles and γ rays ensues. This is an extremely slow process for black holes about the mass of the Sun (produced by supernovas) or larger ones (like those thought to be at galactic centers), taking on the order of 10^{67} years or longer! Smaller black holes would evaporate faster, but they are only speculated to exist as remnants of the Big Bang. Searches for characteristic γ-ray bursts have produced events attributable to more mundane objects like neutron stars accreting matter.

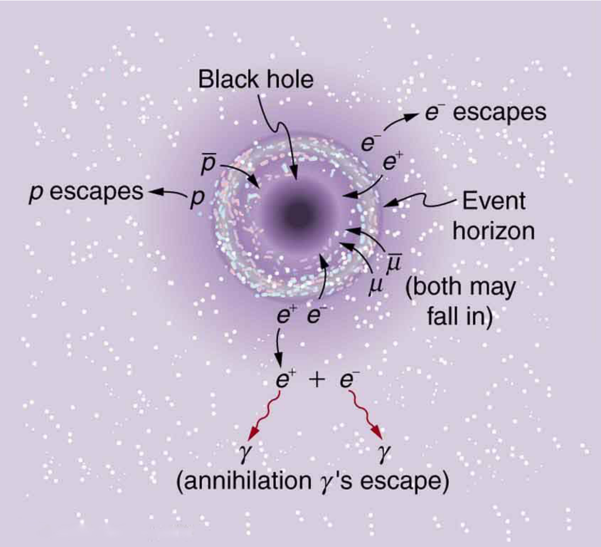


This Hubble Space Telescope photograph shows the extremely energetic core of the NGC 4261 galaxy. With the superior resolution of the orbiting telescope, it has been possible to observe the rotation of an accretion disk around the energy-producing object as well as to map jets of material being ejected from the object. A supermassive black hole is consistent with these observations, but other possibilities are not quite eliminated. (credit: NASA and ESA)



The control room of the LIGO gravitational wave detector. Gravitational waves will cause extremely small vibrations in a mass in this detector, which will be detected by laser interferometer techniques. Such detection in coincidence with other detectors and with astronomical events, such as supernovas, would provide direct evidence of gravitational waves. (credit: Tobin Fricke)

Stephen Hawking (b. 1942) has made many contributions to the theory of quantum gravity. Hawking is a long-time survivor of ALS and has produced popular books on general relativity, cosmology, and quantum gravity. (credit: Lwp Kommunikáció)



Gravity and quantum mechanics

come into play when a black hole creates a particle-antiparticle pair from the energy in its gravitational field. One member of the pair falls into the hole while the other escapes, removing energy and shrinking the black hole. The search is on for the characteristic energy.

Wormholes and time travel

The subject of time travel captures the imagination. Theoretical physicists, such as the American Kip Thorne, have treated the subject seriously, looking into the possibility that falling into a black hole could result in popping up in another time and place—a trip through a so-called wormhole. Time travel and wormholes appear in innumerable science fiction dramatizations, but the consensus is that time travel is not possible in theory. While still debated, it appears that quantum gravity effects inside a black hole prevent time travel due to the creation of particle pairs. Direct evidence is elusive.

The shortest time

Theoretical studies indicate that, at extremely high energies and correspondingly early in the universe, quantum fluctuations may make time intervals meaningful only down to some finite time limit. Early work indicated that this might be the case for times as long as 10^{-43} s, the time at which all forces were unified. If so, then it would be meaningless to consider the universe at times earlier than this. Subsequent studies indicate that the crucial time may be as short as 10^{-95} s. But the point remains—quantum gravity seems to imply that there is no such thing as a vanishingly short time. Time may, in fact, be grainy with no meaning to time intervals shorter than some tiny but finite size.

The future of quantum gravity

Not only is quantum gravity in its infancy, no one knows how to get started on a theory of gravitons and unification of forces. The energies at which TOE should be valid may be so high (at least 10^{19} GeV) and the necessary particle separation so small (less than 10^{-35} m) that only indirect evidence can provide clues. For some time, the common lament of theoretical physicists was one so familiar to struggling students—how do you even get started? But Hawking and others have made a start, and the approach many theorists have taken is called Superstring theory, the topic of the Superstrings.

Section Summary

- Einstein's theory of general relativity includes accelerated frames and, thus, encompasses special relativity and gravity. Created by use of careful thought experiments, it has been repeatedly verified by real experiments.
- One direct result of this behavior of nature is the gravitational lensing of light by massive objects, such as galaxies, also seen in the microlensing of light by smaller bodies in our galaxy.
- Another prediction is the existence of black holes, objects for which the escape velocity is greater than the speed of light and from which nothing can escape.
- The event horizon is the distance from the object at which the escape velocity equals the speed of light *c*. It is called the Schwarzschild radius R_{S} and is given by
  Equation:

  $$R_{\mathrm{S}} = \frac{2\mathrm{GM}}{c^2},$$

  where G is the universal gravitational constant, and M is the mass of the body.
- Physics is unknown inside the event horizon, and the possibility of wormholes and time travel are being studied.
- Candidates for black holes may power the extremely energetic emissions of quasars, distant objects that seem to be early stages of

galactic evolution.
- Neutron stars are stellar remnants, having the density of a nucleus, that hint that black holes could form from supernovas, too.
- Gravitational waves are wrinkles in space, predicted by general relativity but not yet observed, caused by changes in very massive objects.
- Quantum gravity is an incompletely developed theory that strives to include general relativity, quantum mechanics, and unification of forces (thus, a TOE).
- One unconfirmed connection between general relativity and quantum mechanics is the prediction of characteristic radiation from just outside black holes.

Conceptual Questions

Exercise:

Problem:

Quantum gravity, if developed, would be an improvement on both general relativity and quantum mechanics, but more mathematically difficult. Under what circumstances would it be necessary to use quantum gravity? Similarly, under what circumstances could general relativity be used? When could special relativity, quantum mechanics, or classical physics be used?

Exercise:

Problem:

Does observed gravitational lensing correspond to a converging or diverging lens? Explain briefly.

Exercise:

Problem:

Suppose you measure the red shifts of all the images produced by gravitational lensing, such as in [link].You find that the central image has a red shift less than the outer images, and those all have the same red shift. Discuss how this not only shows that the images are of the same object, but also implies that the red shift is not affected by taking different paths through space. Does it imply that cosmological red shifts are not caused by traveling through space (light getting tired, perhaps)?

Exercise:

Problem:

What are gravitational waves, and have they yet been observed either directly or indirectly?

Exercise:

Problem:

Is the event horizon of a black hole the actual physical surface of the object?

Exercise:

Problem:

Suppose black holes radiate their mass away and the lifetime of a black hole created by a supernova is about 10^{67} years. How does this lifetime compare with the accepted age of the universe? Is it surprising that we do not observe the predicted characteristic radiation?

Problems & Exercises

Exercise:

Problem:

What is the Schwarzschild radius of a black hole that has a mass eight times that of our Sun? Note that stars must be more massive than the Sun to form black holes as a result of a supernova.

Solution:

23.6 km

Exercise:

Problem:

Black holes with masses smaller than those formed in supernovas may have been created in the Big Bang. Calculate the radius of one that has a mass equal to the Earth’s.

Exercise:

Problem:

Supermassive black holes are thought to exist at the center of many galaxies.

(a) What is the radius of such an object if it has a mass of 10^9 Suns?

(b) What is this radius in light years?

Solution:

(a) 2.95×10^{12} m

(b) 3.12×10^{-4} ly

Exercise:

Problem: Construct Your Own Problem

Consider a supermassive black hole near the center of a galaxy. Calculate the radius of such an object based on its mass. You must consider how much mass is reasonable for these large objects, and which is now nearly directly observed. (Information on black holes posted on the Web by NASA and other agencies is reliable, for example.)

Glossary

black holes
objects having such large gravitational fields that things can fall in, but nothing, not even light, can escape

general relativity
Einstein’s theory thatdescribes all types of relative motion including accelerated motion and the effects of gravity

gravitational waves
mass-created distortions in space that propagate at the speed of light and that are predicted by general relativity

escape velocity
takeoff velocity when kinetic energy just cancels gravitational potential energy

event horizon
the distance from the object at which the escape velocity is exactly the speed of light

neutron stars
literally a star composed of neutrons

Schwarzschild radius
the radius of the event horizon

thought experiment

mental analysis of certain carefully and clearly defined situations to develop an idea

quasars

the moderately distant galaxies that emit as much or more energy than a normal galaxy

Quantum gravity

the theory that deals with particle exchange of gravitons as the mechanism for the force

Superstrings

- Define Superstring theory.
- Explain the relationship between Superstring theory and the Big Bang.

Introduced earlier in GUTS: The Unification of Forces **Superstring theory** is an attempt to unify gravity with the other three forces and, thus, must contain quantum gravity. The main tenet of Superstring theory is that fundamental particles, including the graviton that carries the gravitational force, act like one-dimensional vibrating strings. Since gravity affects the time and space in which all else exists, Superstring theory is an attempt at a Theory of Everything (TOE). Each independent quantum number is thought of as a separate dimension in some super space (analogous to the fact that the familiar dimensions of space are independent of one another) and is represented by a different type of Superstring. As the universe evolved after the Big Bang and forces became distinct (spontaneous symmetry breaking), some of the dimensions of superspace are imagined to have curled up and become unnoticed.

Forces are expected to be unified only at extremely high energies and at particle separations on the order of 10^{-35} m. This could mean that Superstrings must have dimensions or wavelengths of this size or smaller. Just as quantum gravity may imply that there are no time intervals shorter than some finite value, it also implies that there may be no sizes smaller than some tiny but finite value. That may be about 10^{-35} m. If so, and if Superstring theory can explain all it strives to, then the structures of Superstrings are at the lower limit of the smallest possible size and can have no further substructure. This would be the ultimate answer to the question the ancient Greeks considered. There is a finite lower limit to space.

Not only is Superstring theory in its infancy, it deals with dimensions about 17 orders of magnitude smaller than the 10^{-18} m details that we have been able to observe directly. It is thus relatively unconstrained by experiment, and there are a host of theoretical possibilities to choose from. This has led theorists to make choices subjectively (as always) on what is the most elegant theory, with less hope than usual that experiment will guide them. It has also led to speculation of alternate universes, with their Big Bangs creating each new universe with a random set of rules. These speculations may not be tested even in principle, since an alternate universe is by definition unattainable. It is something like exploring a self-consistent field of mathematics, with its axioms and rules of logic that are not consistent with nature. Such endeavors have often given insight to mathematicians and scientists alike and occasionally have been directly related to the description of new discoveries.

Section Summary

- Superstring theory holds that fundamental particles are one-dimensional vibrations analogous to those on strings and is an attempt at a theory of quantum gravity.

Problems & Exercises

Exercise:

Problem:

The characteristic length of entities in Superstring theory is approximately 10^{-35} m.

(a) Find the energy in GeV of a photon of this wavelength.

(b) Compare this with the average particle energy of 10^{19} GeV needed for unification of forces.

Solution:

(a) 1×10^{20}

(b) 10 times greater

Glossary

Superstring theory

a theory to unify gravity with the other three forces in which the fundamental particles are considered to act like one-dimensional vibrating strings

Dark Matter and Closure

- Discuss the existence of dark matter.
- Explain neutrino oscillations and their consequences.

One of the most exciting problems in physics today is the fact that there is far more matter in the universe than we can see. The motion of stars in galaxies and the motion of galaxies in clusters imply that there is about 10 times as much mass as in the luminous objects we can see. The indirectly observed non-luminous matter is called **dark matter**. Why is dark matter a problem? For one thing, we do not know what it is. It may well be 90% of all matter in the universe, yet there is a possibility that it is of a completely unknown form—a stunning discovery if verified. Dark matter has implications for particle physics. It may be possible that neutrinos actually have small masses or that there are completely unknown types of particles. Dark matter also has implications for cosmology, since there may be enough dark matter to stop the expansion of the universe. That is another problem related to dark matter—we do not know how much there is. We keep finding evidence for more matter in the universe, and we have an idea of how much it would take to eventually stop the expansion of the universe, but whether there is enough is still unknown.

Evidence

The first clues that there is more matter than meets the eye came from the Swiss-born American astronomer Fritz Zwicky in the 1930s; some initial work was also done by the American astronomer Vera Rubin. Zwicky measured the velocities of stars orbiting the galaxy, using the relativistic Doppler shift of their spectra (see [link](a)). He found that velocity varied with distance from the center of the galaxy, as graphed in [link](b). If the mass of the galaxy was concentrated in its center, as are its luminous stars, the velocities should decrease as the square root of the distance from the center. Instead, the velocity curve is almost flat, implying that there is a tremendous amount of matter in the galactic halo. Although not immediately recognized for its significance, such measurements have now been made for many galaxies, with similar results. Further, studies of galactic clusters have also indicated that galaxies have a mass distribution greater than that obtained from their brightness (proportional to the number of stars), which also extends into large halos surrounding the luminous parts of galaxies. Observations of other EM wavelengths, such as radio waves and X rays, have similarly confirmed the existence of dark matter. Take, for example, X rays in the relatively dark space between galaxies, which indicates the presence of previously unobserved hot, ionized gas (see [link] (c)).

Theoretical Yearnings for Closure

Is the universe open or closed? That is, will the universe expand forever or will it stop, perhaps to contract? This, until recently, was a question of whether there is enough gravitation to stop the expansion of the universe. In the past few years, it has become a question of the combination of gravitation and what is called the **cosmological constant**. The cosmological constant was invented by Einstein to prohibit the expansion or contraction of the universe. At the time he developed general relativity, Einstein considered that an illogical possibility. The cosmological constant was discarded after Hubble discovered the expansion, but has been re-invoked in recent years.

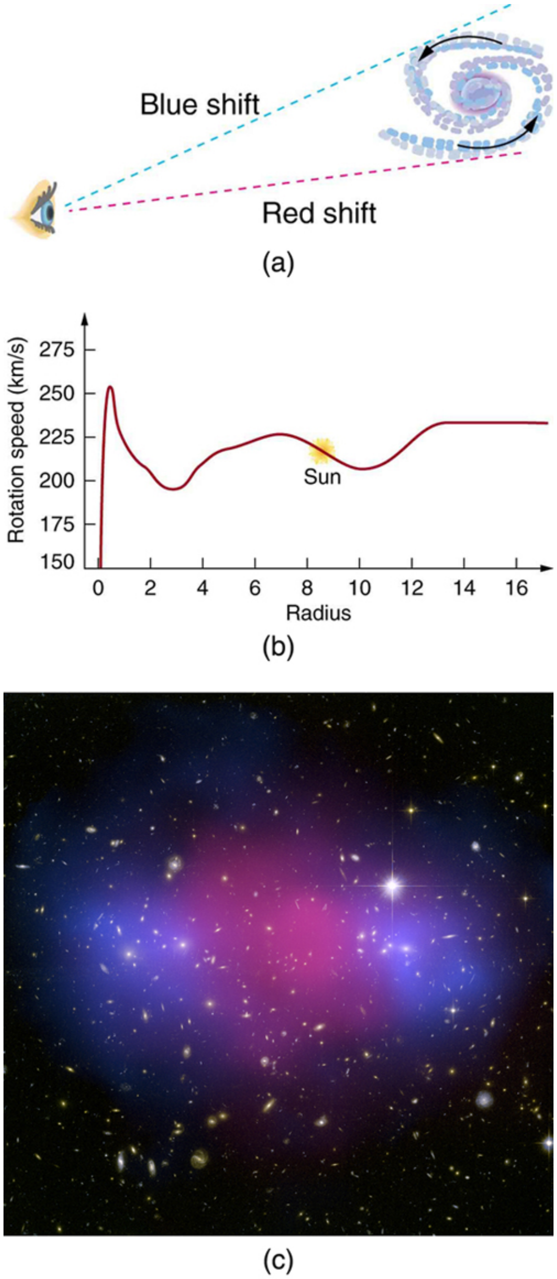
Gravitational attraction between galaxies is slowing the expansion of the universe, but the amount of slowing down is not known directly. In fact, the cosmological constant can counteract gravity’s effect. As recent measurements indicate, the universe is expanding *faster* now than in the past—perhaps a “modern inflationary era” in which the dark energy is thought to be causing the expansion of the present-day universe to accelerate. If the expansion rate were affected by gravity alone, we should be able to see that the expansion rate between distant galaxies was once greater than it is now. However, measurements show it was *less* than now. We can, however, calculate the amount of slowing based on the average density of matter we observe directly. Here we have a definite answer—there is far less visible matter than needed to stop expansion. The **critical density** ρ_c is defined to be the density needed to just halt universal expansion in a universe with no cosmological constant. It is estimated to be about

Equation:

$$\rho_{\text{c}} \approx 10^{-26}\ \text{kg/m}^3.$$

However, this estimate of ρ_{c} is only good to about a factor of two, due to uncertainties in the expansion rate of the universe. The critical density is equivalent to an average of only a few nucleons per cubic meter, remarkably small and indicative of how truly empty intergalactic space is. Luminous matter seems to account for roughly 0.5% to 2% of the critical density, far less than that needed for closure. Taking into account the amount of dark matter we detect indirectly and all other types of indirectly observed normal matter, there is only 10% to 40% of what is needed for closure. If we are able to refine the measurements of expansion rates now and in the past, we will have our answer regarding the curvature of space and we will determine a value for the cosmological constant to justify this observation. Finally, the most recent measurements of the CMBR have implications for the cosmological constant, so it is not simply a device concocted for a single purpose.

After the recent experimental discovery of the cosmological constant, most researchers feel that the universe should be just barely open. Since matter can be thought to curve the space around it, we call an open universe **negatively curved**. This means that you can in principle travel an unlimited distance in any direction. A universe that is closed is called **positively curved**. This means that if you travel far enough in any direction, you will return to your starting point, analogous to circumnavigating the Earth. In between these two is a **flat (zero curvature) universe**. The recent discovery of the cosmological constant has shown the universe is very close to flat, and will expand forever. Why do theorists feel the universe is flat? Flatness is a part of the inflationary scenario that helps explain the flatness of the microwave background. In fact, since general relativity implies that matter creates the space in which it exists, there is a special symmetry to a flat universe.



Evidence for dark matter: (a) We can measure the velocities of stars relative to their galaxies by observing the Doppler shift in emitted light, usually using the hydrogen spectrum. These measurements indicate the rotation of a spiral galaxy. (b) A graph of velocity versus distance from the galactic center shows that the velocity does not decrease as it would if the matter were concentrated in luminous stars. The flatness of the curve implies a massive galactic halo of dark matter extending beyond the visible stars. (c) This is a computer-generated image of X rays from a galactic cluster. The X rays indicate the presence of otherwise unseen hot clouds of ionized gas in the regions of space previously considered more empty. (credit: NASA, ESA, CXC, M. Bradac (University of California, Santa Barbara), and S. Allen (Stanford University))

What Is the Dark Matter We See Indirectly?

There is no doubt that dark matter exists, but its form and the amount in existence are two facts that are still being studied vigorously. As always, we seek to explain new observations in terms of known principles. However, as more discoveries are made, it is becoming more and more difficult to explain dark matter as a known type of matter.

One of the possibilities for normal matter is being explored using the Hubble Space Telescope and employing the lensing effect of gravity on light (see [link]). Stars glow because of nuclear fusion in them, but planets are visible primarily by reflected light. Jupiter, for example, is too small to ignite fusion in its core and become a star, but we can see sunlight reflected from it, since we are relatively close. If Jupiter orbited another star, we would not be able to see it directly. The question is open as to how many planets or other bodies smaller than about 1/1000 the mass of the Sun are there. If such bodies pass between us and a star, they will not block the star's light, being too small, but they will form a gravitational lens, as discussed in General Relativity and Quantum Gravity.

In a process called **microlensing**, light from the star is focused and the star appears to brighten in a characteristic manner. Searches for dark matter in this form are particularly interested in galactic halos because of the huge amount of mass that seems to be there. Such microlensing objects are thus called **massive compact halo objects**, or **MACHOs**. To date, a few MACHOs have been observed, but not predominantly in galactic halos, nor in the numbers needed to explain dark matter.

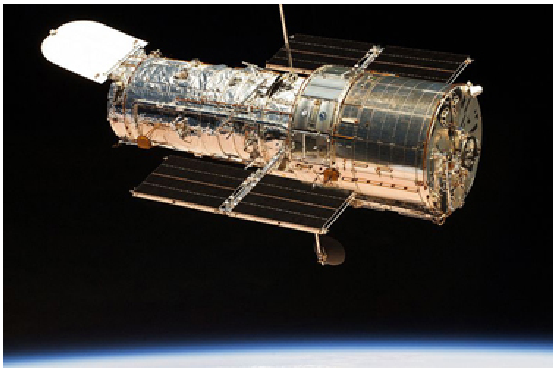
MACHOs are among the most conventional of unseen objects proposed to explain dark matter. Others being actively pursued are red dwarfs, which are small dim stars, but too few have been seen so far, even with the Hubble Telescope, to be of significance. Old remnants of stars called white dwarfs are also under consideration, since they contain about a solar mass, but are small as the Earth and may dim to the point that we ordinarily do not observe them. While white dwarfs are known, old dim ones are not. Yet another possibility is the existence of large numbers of smaller than stellar mass black holes left from the Big Bang—here evidence is entirely absent.

There is a very real possibility that dark matter is composed of the known neutrinos, which may have small, but finite, masses. As discussed earlier, neutrinos are thought to be massless, but we only have upper limits on their masses, rather than knowing they are exactly zero. So far, these upper limits come from difficult measurements of total energy emitted in the decays and reactions in which neutrinos are involved. There is an amusing possibility of proving that neutrinos have mass in a completely different way.

We have noted in Particles, Patterns, and Conservation Laws that there are three flavors of neutrinos (ν_e, v_μ, and v_τ) and that the weak interaction could change quark flavor. It should also change neutrino flavor—that is, any type of neutrino could change spontaneously into any other, a process called **neutrino oscillations**. However, this can occur only if neutrinos have a mass. Why? Crudely, because if neutrinos are massless, they must travel at the speed of light and time will not pass for them, so that they cannot change without an interaction. In 1999, results began to be published containing convincing evidence that neutrino oscillations do occur. Using the Super-Kamiokande detector in Japan, the oscillations have been observed and are being verified and further explored at present at the same facility and others.

Neutrino oscillations may also explain the low number of observed solar neutrinos. Detectors for observing solar neutrinos are specifically designed to detect electron neutrinos ν_e produced in huge numbers by fusion in the Sun. A large fraction of electron neutrinos ν_e may be changing flavor to muon neutrinos v_μ on their way out of the Sun, possibly enhanced by specific interactions, reducing the flux of electron neutrinos to observed levels. There is also a discrepancy in observations of neutrinos produced in cosmic ray showers. While these showers of radiation produced by extremely energetic cosmic rays should contain twice as many v_μ s as ν_e s, their numbers are nearly equal. This may be explained by neutrino oscillations from muon flavor to electron flavor. Massive neutrinos are a particularly appealing possibility for explaining dark matter, since their existence is consistent with a large body of known information and explains more than dark matter. The question is not settled at this writing.

The most radical proposal to explain dark matter is that it consists of previously unknown leptons (sometimes obtusely referred to as non-baryonic matter). These are called **weakly interacting massive particles**, or **WIMPs**, and would also be chargeless, thus interacting negligibly with normal matter, except through gravitation. One proposed group of WIMPs would have masses several orders of magnitude greater than nucleons and are sometimes called **neutralinos**. Others are called **axions** and would have masses about 10^{-10} that of an electron mass. Both neutralinos and axions would be gravitationally attached to galaxies, but because they are chargeless and only feel the weak force, they would be in a halo rather than interact and coalesce into spirals, and so on, like normal matter (see [link]).



The Hubble Space Telescope is producing exciting data with its corrected optics and with the absence of atmospheric distortion. It has observed some MACHOs, disks of material around stars thought to precede planet formation, black hole candidates, and collisions of comets with Jupiter. (credit: NASA (crew of STS-125))

Dark matter may shepherd normal matter gravitationally in space, as this stream moves the leaves. Dark matter may be invisible and even move through the normal matter, as neutrinos penetrate us without small-scale effect. (credit: Shinichi Sugiyama)

Some particle theorists have built WIMPs into their unified force theories and into the inflationary scenario of the evolution of the universe so popular today. These particles would have been produced in just the correct numbers to make the universe flat, shortly after the Big Bang. The proposal is radical in the sense that it invokes entirely new forms of matter, in fact *two* entirely new forms, in order to explain dark matter and other phenomena. WIMPs have the extra burden of automatically being very difficult to observe directly. This is somewhat analogous to quark confinement, which guarantees that quarks are there, but they can never be seen directly. One of the primary goals of the LHC at CERN, however, is to produce and detect WIMPs. At any rate, before WIMPs are accepted as the best explanation, all other possibilities utilizing known phenomena will have to be shown inferior. Should that occur, we will be in the unanticipated position of admitting that, to date, all we know is only 10% of what exists.

A far cry from the days when people firmly believed themselves to be not only the center of the universe, but also the reason for its existence.

Section Summary

- Dark matter is non-luminous matter detected in and around galaxies and galactic clusters.
- It may be 10 times the mass of the luminous matter in the universe, and its amount may determine whether the universe is open or closed (expands forever or eventually stops).
- The determining factor is the critical density of the universe and the cosmological constant, a theoretical construct intimately related to the expansion and closure of the universe.
- The critical density ρ_c is the density needed to just halt universal expansion. It is estimated to be approximately 10^{-26} kg/m^3.
- An open universe is negatively curved, a closed universe is positively curved, whereas a universe with exactly the critical density is flat.
- Dark matter's composition is a major mystery, but it may be due to the suspected mass of neutrinos or a completely unknown type of leptonic matter.
- If neutrinos have mass, they will change families, a process known as neutrino oscillations, for which there is growing evidence.

Conceptual Questions

Exercise:

Problem:

Discuss the possibility that star velocities at the edges of galaxies being greater than expected is due to unknown properties of gravity rather than to the existence of dark matter. Would this mean, for example, that gravity is greater or smaller than expected at large distances? Are there other tests that could be made of gravity at large distances, such as observing the motions of neighboring galaxies?

Exercise:

Problem:

How does relativistic time dilation prohibit neutrino oscillations if they are massless?

Exercise:

Problem:

If neutrino oscillations do occur, will they violate conservation of the various lepton family numbers (L_e, L_μ, and L_τ)? Will neutrino oscillations violate conservation of the total number of leptons?

Exercise:

Problem:

Lacking direct evidence of WIMPs as dark matter, why must we eliminate all other possible explanations based on the known forms of matter before we invoke their existence?

Problems Exercises

Exercise:

Problem:

If the dark matter in the Milky Way were composed entirely of MACHOs (evidence shows it is not), approximately how many would there have to be? Assume the average mass of a MACHO is 1/1000 that of the Sun, and that dark matter has a mass 10 times that of the luminous Milky Way galaxy with its 10^{11} stars of average mass 1.5 times the Sun's mass.

Solution:
Equation:

$$1.5 \times 10^{15}$$

Exercise:

Problem:

The critical mass density needed to just halt the expansion of the universe is approximately $10^{-26}\ \mathrm{kg/m^3}$.

(a) Convert this to $\mathrm{eV}/c^2 \cdot \mathrm{m}^3$.

(b) Find the number of neutrinos per cubic meter needed to close the universe if their average mass is $7\ \mathrm{eV}/c^2$ and they have negligible kinetic energies.

Exercise:

Problem:

Assume the average density of the universe is 0.1 of the critical density needed for closure. What is the average number of protons per cubic meter, assuming the universe is composed mostly of hydrogen?

Solution:
Equation:

$$0.6\ \mathrm{m}^{-3}$$

Exercise:

Problem:

To get an idea of how empty deep space is on the average, perform the following calculations:

(a) Find the volume our Sun would occupy if it had an average density equal to the critical density of $10^{-26}\ \mathrm{kg/m^3}$ thought necessary to halt the expansion of the universe.

(b) Find the radius of a sphere of this volume in light years.

(c) What would this radius be if the density were that of luminous matter, which is approximately 5% that of the critical density?

(d) Compare the radius found in part (c) with the 4-ly average separation of stars in the arms of the Milky Way.

Glossary

axions
: a type of WIMPs having masses about 10^{-10} of an electron mass

cosmological constant
: a theoretical construct intimately related to the expansion and closure of the universe

critical density
: the density of matter needed to just halt universal expansion

dark matter
: indirectly observed non-luminous matter

flat (zero curvature) universe
: a universe that is infinite but not curved

microlensing
: a process in which light from a distant star is focused and the star appears to brighten in a characteristic manner, when a small body (smaller than about 1/1000 the mass of the Sun) passes between us and the star

MACHOs
: massive compact halo objects; microlensing objects of huge mass

neutrino oscillations
: a process in which any type of neutrino could change spontaneously into any other

neutralinos
a type of WIMPs having masses several orders of magnitude greater than nucleon masses

negatively curved
an open universe that expands forever

positively curved
a universe that is closed and eventually contracts

WIMPs
weakly interacting massive particles; chargeless leptons (non-baryonic matter) interacting negligibly with normal matter

Complexity and Chaos

- Explain complex systems.
- Discuss chaotic behavior of different systems.

Much of what impresses us about physics is related to the underlying connections and basic simplicity of the laws we have discovered. The language of physics is precise and well defined because many basic systems we study are simple enough that we can perform controlled experiments and discover unambiguous relationships. Our most spectacular successes, such as the prediction of previously unobserved particles, come from the simple underlying patterns we have been able to recognize. But there are systems of interest to physicists that are inherently complex. The simple laws of physics apply, of course, but complex systems may reveal patterns that simple systems do not. The emerging field of **complexity** is devoted to the study of complex systems, including those outside the traditional bounds of physics. Of particular interest is the ability of complex systems to adapt and evolve.

What are some examples of complex adaptive systems? One is the primordial ocean. When the oceans first formed, they were a random mix of elements and compounds that obeyed the laws of physics and chemistry. In a relatively short geological time (about 500 million years), life had emerged. Laboratory simulations indicate that the emergence of life was far too fast to have come from random combinations of compounds, even if driven by lightning and heat. There must be an underlying ability of the complex system to organize itself, resulting in the self-replication we recognize as life. Living entities, even at the unicellular level, are highly organized and systematic. Systems of living organisms are themselves complex adaptive systems. The grandest of these evolved into the biological system we have today, leaving traces in the geological record of steps taken along the way.

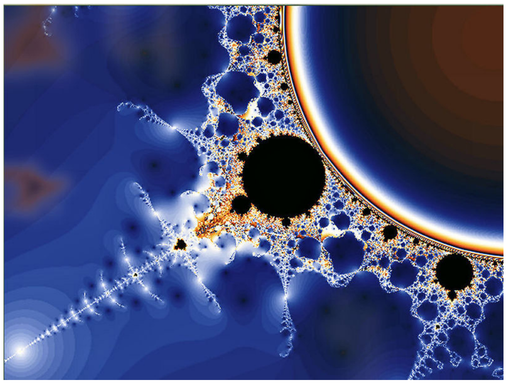
Complexity as a discipline examines complex systems, how they adapt and evolve, looking for similarities with other complex adaptive systems. Can, for example, parallels be drawn between biological evolution and the evolution of *economic systems*? Economic systems do emerge quickly, they show tendencies for self-organization, they are complex (in the number and types of transactions), and they adapt and evolve. Biological systems do all the same types of things. There are other examples of complex adaptive systems being studied for fundamental similarities. *Cultures* show signs of adaptation and evolution. The comparison of different cultural evolutions may bear fruit as well as comparisons to biological evolution. *Science* also is a complex system of human interactions, like culture and economics, that adapts to new information and political pressure, and evolves, usually becoming more organized rather than less. Those who study *creative thinking* also see parallels with complex systems. Humans sometimes organize almost random pieces of information, often subconsciously while doing other things, and come up with brilliant creative insights. The development of *language* is another complex adaptive system that may show similar tendencies. *Artificial intelligence* is an overt attempt to devise an adaptive system that will self-organize and evolve in the same manner as an intelligent living being learns. These are a few of the broad range of topics being studied by those who investigate complexity. There are now institutes, journals, and meetings, as well as popularizations of the emerging topic of complexity.

In traditional physics, the discipline of complexity may yield insights in certain areas. Thermodynamics treats systems on the average, while statistical mechanics deals in some detail with complex systems of atoms and molecules in random thermal motion. Yet there is organization, adaptation, and evolution in those complex systems. Non-equilibrium phenomena, such as heat transfer and phase changes, are characteristically complex in detail, and new approaches to them may evolve from complexity as a discipline. Crystal growth is another example of self-organization spontaneously emerging in a complex system. Alloys are also inherently complex mixtures that show certain simple characteristics implying some self-organization. The organization of iron atoms into magnetic domains as they cool is another. Perhaps insights into these difficult areas will emerge from complexity. But at the minimum, the discipline of complexity is another example of human effort to understand and organize the universe around us, partly rooted in the discipline of physics.

A predecessor to complexity is the topic of chaos, which has been widely publicized and has become a discipline of its own. It is also based partly in physics and treats broad classes of phenomena from many disciplines. **Chaos** is a word used to describe systems whose outcomes are extremely sensitive to initial conditions. The orbit of the planet Pluto, for example, may be chaotic in that it can change tremendously due to small interactions with other planets. This makes its long-term behavior impossible to predict with precision, just as we cannot tell precisely where a decaying Earth satellite will land or how many pieces it will break into. But the discipline of chaos has found ways to deal with such systems and has been applied to apparently unrelated systems. For example, the heartbeat of people with certain types of potentially lethal arrhythmias seems to be chaotic, and this knowledge may allow more sophisticated monitoring and recognition of the need for intervention.

Chaos is related to complexity. Some chaotic systems are also inherently complex; for example, vortices in a fluid as opposed to a double pendulum. Both are chaotic and not predictable in the same sense as other systems. But there can be organization in chaos and it can also be quantified. Examples of chaotic systems are beautiful fractal patterns such as in [link]. Some chaotic systems exhibit self-organization, a type of stable chaos. The orbits of the planets in our solar system, for example, may be chaotic (we are not certain yet). But they are definitely organized and systematic, with a simple formula describing the orbital radii of the first eight planets *and* the asteroid belt. Large-scale vortices in Jupiter's atmosphere are chaotic, but the Great Red Spot is a stable self-organization of rotational energy. (See [link].) The Great Red Spot has been in existence for at least 400 years and is a complex self-adaptive system.

The emerging field of complexity, like the now almost traditional field of chaos, is partly rooted in physics. Both attempt to see similar systematics in a very broad range of phenomena and, hence, generate a better understanding of them. Time will tell what impact these fields have on more traditional areas of physics as well as on the other disciplines they relate to.



This image is related to the Mandelbrot set, a complex mathematical form that is chaotic. The patterns are infinitely fine as you look closer and closer, and they indicate order in the presence of chaos. (credit: Gilberto Santa Rosa)

The Great Red Spot on Jupiter is an example of self-organization in a complex and chaotic system. Smaller vortices in Jupiter's atmosphere behave chaotically, but the triple-Earth-size spot is self-organized and stable for at least hundreds of years. (credit: NASA)

Section Summary

- Complexity is an emerging field, rooted primarily in physics, that considers complex adaptive systems and their evolution, including self-organization.
- Complexity has applications in physics and many other disciplines, such as biological evolution.
- Chaos is a field that studies systems whose properties depend extremely sensitively on some variables and whose evolution is impossible to predict.
- Chaotic systems may be simple or complex.
- Studies of chaos have led to methods for understanding and predicting certain chaotic behaviors.

Conceptual Questions

Exercise:

Problem:

Must a complex system be adaptive to be of interest in the field of complexity? Give an example to support your answer.

Exercise:

Problem: State a necessary condition for a system to be chaotic.

Glossary

complexity

an emerging field devoted to the study of complex systems

chaos

word used to describe systems the outcomes of which are extremely sensitive to initial conditions

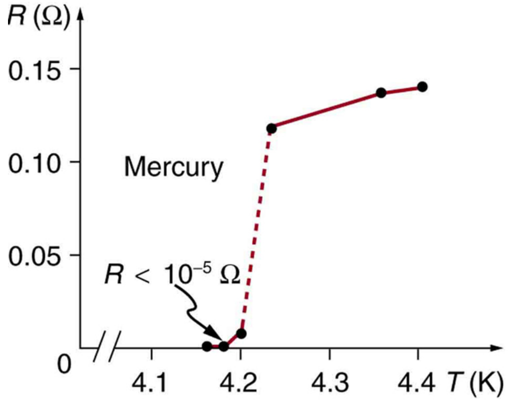
High-temperature Superconductors

- Identify superconductors and their uses.
- Discuss the need for a high-T_c superconductor.

Superconductors are materials with a resistivity of zero. They are familiar to the general public because of their practical applications and have been mentioned at a number of points in the text. Because the resistance of a piece of superconductor is zero, there are no heat losses for currents through them; they are used in magnets needing high currents, such as in MRI machines, and could cut energy losses in power transmission. But most superconductors must be cooled to temperatures only a few kelvin above absolute zero, a costly procedure limiting their practical applications. In the past decade, tremendous advances have been made in producing materials that become superconductors at relatively high temperatures. There is hope that room temperature superconductors may someday be manufactured.

Superconductivity was discovered accidentally in 1911 by the Dutch physicist H. Kamerlingh Onnes (1853–1926) when he used liquid helium to cool mercury. Onnes had been the first person to liquefy helium a few years earlier and was surprised to observe the resistivity of a mediocre conductor like mercury drop to zero at a temperature of 4.2 K. We define the temperature at which and below which a material becomes a superconductor to be its **critical temperature**, denoted by T_c. (See [link].) Progress in understanding how and why a material became a superconductor was relatively slow, with the first workable theory coming in 1957. Certain other elements were also found to become superconductors, but all had T_c s less than 10 K, which are expensive to maintain. Although Onnes received a Nobel prize in 1913, it was primarily for his work with liquid helium.

In 1986, a breakthrough was announced—a ceramic compound was found to have an unprecedented T_c of 35 K. It looked as if much higher critical temperatures could be possible, and by early 1988 another ceramic (this of thallium, calcium, barium, copper, and oxygen) had been found to have $T_c = 125\text{ K}$ (see [link].) The economic potential of perfect conductors saving electric energy is immense for T_c s above 77 K, since that is the temperature of liquid nitrogen. Although liquid helium has a boiling point of 4 K and can be used to make materials superconducting, it costs about $5 per liter. Liquid nitrogen boils at 77 K, but only costs about $0.30 per liter. There was general euphoria at the discovery of these complex ceramic superconductors, but this soon subsided with the sobering difficulty of forming them into usable wires. The first commercial use of a high temperature superconductor is in an electronic filter for cellular phones. High-temperature superconductors are used in experimental apparatus, and they are actively being researched, particularly in thin film applications.

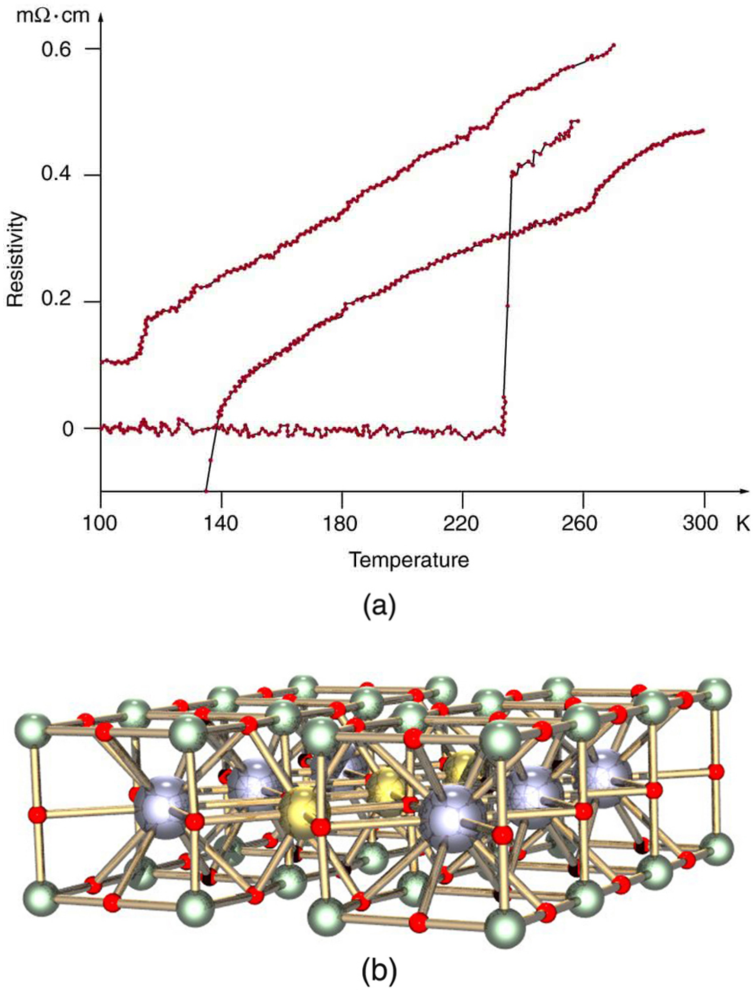


A graph of resistivity versus temperature for a superconductor shows a sharp transition to zero at the critical temperature T_c. High temperature superconductors have verifiable T_c s greater than 125 K, well above the easily achieved 77-K temperature of liquid nitrogen.

One characteristic of a superconductor is that it excludes magnetic flux and, thus, repels other magnets. The small magnet levitated above a high-temperature superconductor, which is cooled by liquid nitrogen, gives evidence that the material is superconducting. When the material warms and becomes conducting, magnetic flux can penetrate it, and the magnet will rest upon it. (credit: Saperaud)

The search is on for even higher T_c superconductors, many of complex and exotic copper oxide ceramics, sometimes including strontium, mercury, or yttrium as well as barium, calcium, and other elements. Room temperature (about 293 K) would be ideal, but any temperature close to room temperature is relatively cheap to produce and maintain. There are persistent reports of T_c s over 200 K and some in the vicinity of 270 K. Unfortunately, these observations are not routinely reproducible, with samples losing their superconducting nature once heated and recooled (cycled) a few times (see [link].) They are now called USOs or unidentified superconducting objects, out of frustration and the refusal of some samples to show high T_c even though produced in the same manner as others. Reproducibility is crucial to discovery, and researchers are justifiably reluctant to claim the breakthrough they all seek. Time will tell whether USOs are real or an experimental quirk.

The theory of ordinary superconductors is difficult, involving quantum effects for widely separated electrons traveling through a material. Electrons couple in a manner that allows them to get through the material without losing energy to it, making it a superconductor. High- T_c superconductors are more difficult to understand theoretically, but theorists seem to be closing in on a workable theory. The difficulty of understanding how electrons can sneak through materials without losing energy in collisions is even greater at higher temperatures, where vibrating atoms should get in the way. Discoverers of high T_c may feel something analogous to what a politician once said upon an unexpected election victory—“I wonder what we did right?”



(a) This graph, adapted from an article in *Physics Today*, shows the behavior of a single sample of a high-temperature superconductor in three different trials. In one case the sample exhibited a T_c of about 230 K, whereas in the others it did not become superconducting at all. The lack of reproducibility is typical of forefront experiments and prohibits definitive conclusions. (b) This colorful diagram shows the complex but systematic nature of the lattice structure of a high-temperature superconducting ceramic. (credit: en:Cadmium, Wikimedia Commons)

Section Summary

- High-temperature superconductors are materials that become superconducting at temperatures well above a few kelvin.
- The critical temperature T_c is the temperature below which a material is superconducting.
- Some high-temperature superconductors have verified T_c s above 125 K, and there are reports of T_c s as high as 250 K.

Conceptual Questions

Exercise:

Problem:

What is critical temperature T_c? Do all materials have a critical temperature? Explain why or why not.

Exercise:

Problem:

Explain how good thermal contact with liquid nitrogen can keep objects at a temperature of 77 K (liquid nitrogen’s boiling point at atmospheric pressure).

Exercise:

Problem:

Not only is liquid nitrogen a cheaper coolant than liquid helium, its boiling point is higher (77 K vs. 4.2 K). How does higher temperature help lower the cost of cooling a material? Explain in terms of the rate of heat transfer being related to the temperature difference between the sample and its surroundings.

Problem Exercises

Exercise:

Problem:

A section of superconducting wire carries a current of 100 A and requires 1.00 L of liquid nitrogen per hour to keep it below its critical temperature. For it to be economically advantageous to use a superconducting wire, the cost of cooling the wire must be less than the cost of energy lost to heat in the wire. Assume that the cost of liquid nitrogen is \$0.30 per liter, and that electric energy costs \$0.10 per kW·h. What is the resistance of a normal wire that costs as much in wasted electric energy as the cost of liquid nitrogen for the superconductor?

Solution:
Equation:

$$0.30\ \Omega$$

Glossary

Superconductors
materials with resistivity of zero

critical temperature
the temperature at which and below which a material becomes a superconductor

Some Questions We Know to Ask

- Identify sample questions to be asked on the largest scales.
- Identify sample questions to be asked on the intermediate scale.
- Identify sample questions to be asked on the smallest scales.

Throughout the text we have noted how essential it is to be curious and to ask questions in order to first understand what is known, and then to go a little farther. Some questions may go unanswered for centuries; others may not have answers, but some bear delicious fruit. Part of discovery is knowing which questions to ask. You have to know something before you can even phrase a decent question. As you may have noticed, the mere act of asking a question can give you the answer. The following questions are a sample of those physicists now know to ask and are representative of the forefronts of physics. Although these questions are important, they will be replaced by others if answers are found to them. The fun continues.

On the Largest Scale

1. *Is the universe open or closed*? Theorists would like it to be just barely closed and evidence is building toward that conclusion. Recent measurements in the expansion rate of the universe and in CMBR support a flat universe. There is a connection to small-scale physics in the type and number of particles that may contribute to closing the universe.
2. *What is dark matter*? It is definitely there, but we really do not know what it is. Conventional possibilities are being ruled out, but one of them still may explain it. The answer could reveal whole new realms of physics and the disturbing possibility that most of what is out there is unknown to us, a completely different form of matter.
3. *How do galaxies form*? They exist since very early in the evolution of the universe and it remains difficult to understand how they evolved so quickly. The recent finer measurements of fluctuations in the CMBR may yet allow us to explain galaxy formation.
4. *What is the nature of various-mass black holes*? Only recently have we become confident that many black hole candidates cannot be explained by other, less exotic possibilities. But we still do not know much about

how they form, what their role in the history of galactic evolution has been, and the nature of space in their vicinity. However, so many black holes are now known that correlations between black hole mass and galactic nuclei characteristics are being studied.

5. *What is the mechanism for the energy output of quasars*? These distant and extraordinarily energetic objects now seem to be early stages of galactic evolution with a supermassive black-hole-devouring material. Connections are now being made with galaxies having energetic cores, and there is evidence consistent with less consuming, supermassive black holes at the center of older galaxies. New instruments are allowing us to see deeper into our own galaxy for evidence of our own massive black hole.
6. *Where do the γ bursts come from*? We see bursts of γ rays coming from all directions in space, indicating the sources are very distant objects rather than something associated with our own galaxy. Some γ bursts finally are being correlated with known sources so that the possibility they may originate in binary neutron star interactions or black holes eating a companion neutron star can be explored.

On the Intermediate Scale

1. *How do phase transitions take place on the microscopic scale*? We know a lot about phase transitions, such as water freezing, but the details of how they occur molecule by molecule are not well understood. Similar questions about specific heat a century ago led to early quantum mechanics. It is also an example of a complex adaptive system that may yield insights into other self-organizing systems.
2. *Is there a way to deal with nonlinear phenomena that reveals underlying connections*? Nonlinear phenomena lack a direct or linear proportionality that makes analysis and understanding a little easier. There are implications for nonlinear optics and broader topics such as chaos.
3. *How do high- T_c superconductors become resistanceless at such high temperatures*? Understanding how they work may help make them more practical or may result in surprises as unexpected as the discovery of superconductivity itself.

4. *There are magnetic effects in materials we do not understand—how do they work*? Although beyond the scope of this text, there is a great deal to learn in condensed matter physics (the physics of solids and liquids). We may find surprises analogous to lasing, the quantum Hall effect, and the quantization of magnetic flux. Complexity may play a role here, too.

On the Smallest Scale

1. *Are quarks and leptons fundamental, or do they have a substructure*? The higher energy accelerators that are just completed or being constructed may supply some answers, but there will also be input from cosmology and other systematics.
2. *Why do leptons have integral charge while quarks have fractional charge*? If both are fundamental and analogous as thought, this question deserves an answer. It is obviously related to the previous question.
3. *Why are there three families of quarks and leptons*? First, does this imply some relationship? Second, why three and only three families?
4. *Are all forces truly equal (unified) under certain circumstances*? They don’t have to be equal just because we want them to be. The answer may have to be indirectly obtained because of the extreme energy at which we think they are unified.
5. *Are there other fundamental forces*? There was a flurry of activity with claims of a fifth and even a sixth force a few years ago. Interest has subsided, since those forces have not been detected consistently. Moreover, the proposed forces have strengths similar to gravity, making them extraordinarily difficult to detect in the presence of stronger forces. But the question remains; and if there are no other forces, we need to ask why only four and why these four.
6. *Is the proton stable*? We have discussed this in some detail, but the question is related to fundamental aspects of the unification of forces. We may never know from experiment that the proton is stable, only that it is very long lived.
7. *Are there magnetic monopoles*? Many particle theories call for very massive individual north- and south-pole particles—magnetic

monopoles. If they exist, why are they so different in mass and elusiveness from electric charges, and if they do not exist, why not?

8. *Do neutrinos have mass*? Definitive evidence has emerged for neutrinos having mass. The implications are significant, as discussed in this chapter. There are effects on the closure of the universe and on the patterns in particle physics.
9. *What are the systematic characteristics of high- Z nuclei*? All elements with $Z = 118$ or less (with the exception of 115 and 117) have now been discovered. It has long been conjectured that there may be an island of relative stability near $Z = 114$, and the study of the most recently discovered nuclei will contribute to our understanding of nuclear forces.

These lists of questions are not meant to be complete or consistently important—you can no doubt add to it yourself. There are also important questions in topics not broached in this text, such as certain particle symmetries, that are of current interest to physicists. Hopefully, the point is clear that no matter how much we learn, there always seems to be more to know. Although we are fortunate to have the hard-won wisdom of those who preceded us, we can look forward to new enlightenment, undoubtedly sprinkled with surprise.

Section Summary

- On the largest scale, the questions which can be asked may be about dark matter, dark energy, black holes, quasars, and other aspects of the universe.
- On the intermediate scale, we can query about gravity, phase transitions, nonlinear phenomena, high- T_c superconductors, and magnetic effects on materials.
- On the smallest scale, questions may be about quarks and leptons, fundamental forces, stability of protons, and existence of monopoles.

Conceptual Questions

Exercise:

Problem:

For experimental evidence, particularly of previously unobserved phenomena, to be taken seriously it must be reproducible or of sufficiently high quality that a single observation is meaningful. Supernova 1987A is not reproducible. How do we know observations of it were valid? The fifth force is not broadly accepted. Is this due to lack of reproducibility or poor-quality experiments (or both)? Discuss why forefront experiments are more subject to observational problems than those involving established phenomena.

Exercise:

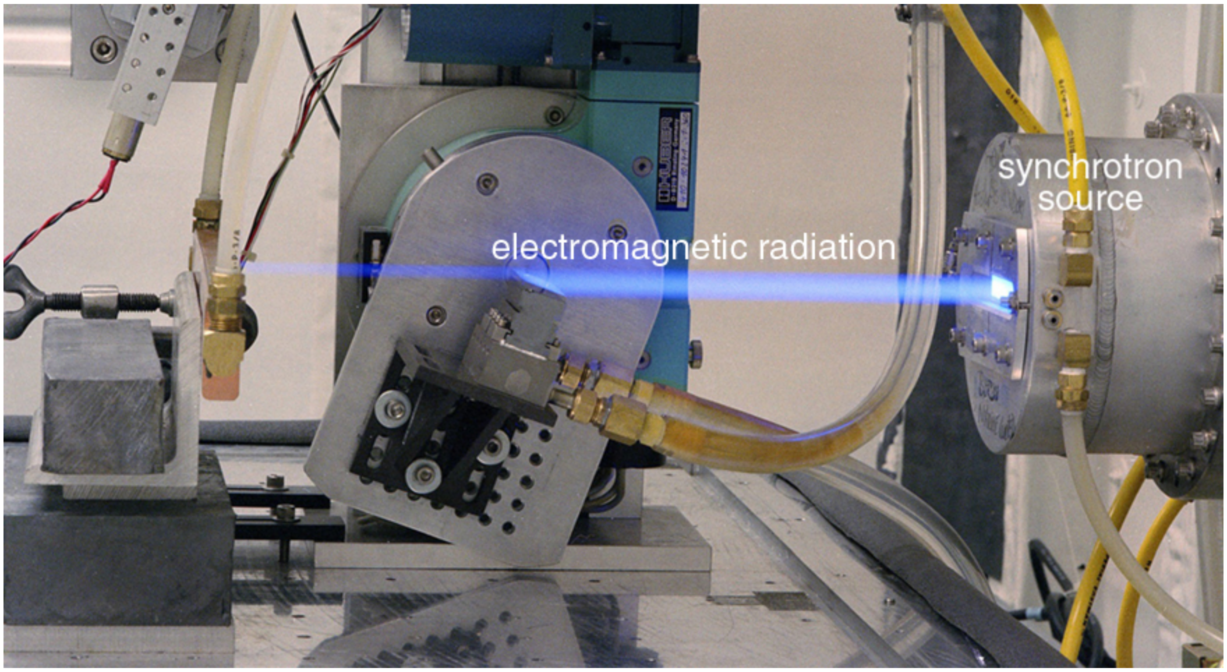
Problem:

Discuss whether you think there are limits to what humans can understand about the laws of physics. Support your arguments.

Introduction to Radioactivity and Nuclear Physics class="introduction"

- Define radioactivity.

The synchrotron source produces electromagnetic radiation, as evident from the visible glow. (credit: United States Department of Energy, via Wikimedia Commons)

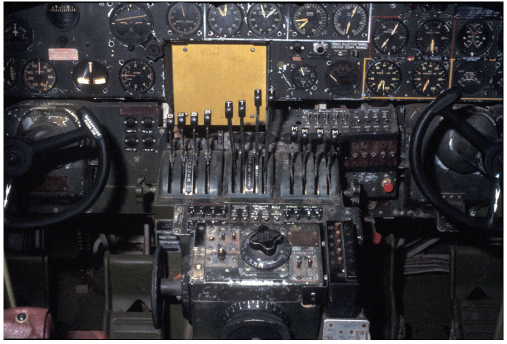


There is an ongoing quest to find substructures of matter. At one time, it was thought that atoms would be the ultimate substructure, but just when the first direct evidence of atoms was obtained, it became clear that they have a substructure and a tiny *nucleus*. The nucleus itself has spectacular characteristics. For example, certain nuclei are unstable, and their decay emits radiations with energies millions of times greater than atomic energies. Some of the mysteries of nature, such as why the core of the earth remains molten and how the sun produces its energy, are explained by nuclear phenomena. The exploration of *radioactivity* and the nucleus revealed fundamental and previously unknown particles, forces, and conservation laws. That exploration has evolved into a search for further underlying structures, such as quarks. In this chapter, the fundamentals of nuclear radioactivity and the nucleus are explored. The following two chapters explore the more important applications of nuclear physics in the field of medicine. We will also explore the basics of what we know about quarks and other substructures smaller than nuclei.

Nuclear Radioactivity

- Explain nuclear radiation.
- Explain the types of radiation—alpha emission, beta emission, and gamma emission.
- Explain the ionization of radiation in an atom.
- Define the range of radiation.

The discovery and study of nuclear radioactivity quickly revealed evidence of revolutionary new physics. In addition, uses for nuclear radiation also emerged quickly—for example, people such as Ernest Rutherford used it to determine the size of the nucleus and devices were painted with radon-doped paint to make them glow in the dark (see [link]). We therefore begin our study of nuclear physics with the discovery and basic features of nuclear radioactivity.



The dials of this World War II aircraft glow in the dark, because they are painted with radium-doped phosphorescent paint. It is a poignant reminder of the dual nature of radiation. Although radium paint dials are conveniently visible day and night, they emit radon, a radioactive gas that is hazardous and is not directly sensed. (credit: U.S. Air Force Photo)

Discovery of Nuclear Radioactivity

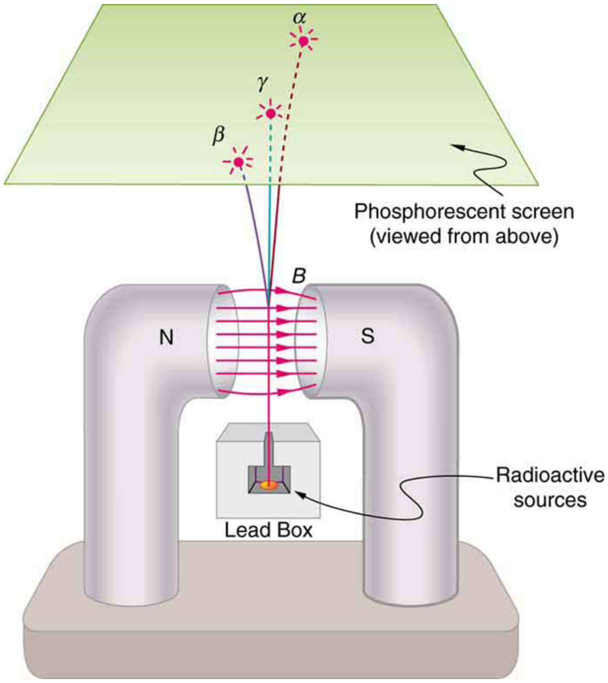
In 1896, the French physicist Antoine Henri Becquerel (1852–1908) accidentally found that a uranium-rich mineral called pitchblende emits invisible, penetrating rays that can darken a photographic plate enclosed in an opaque envelope. The rays therefore carry energy; but amazingly, the pitchblende emits them continuously without any energy input. This is an apparent violation of the law of conservation of energy, one that we now understand is due to the conversion of a small amount of mass into energy, as related in Einstein’s famous equation $E = mc^2$. It was soon evident that Becquerel’s rays originate in the nuclei of the atoms and have other unique characteristics. The emission of these rays is called **nuclear radioactivity** or simply **radioactivity**. The rays themselves are called **nuclear radiation**. A nucleus that spontaneously destroys part of its mass to emit radiation is said to **decay** (a term also used to describe the emission of radiation by atoms in excited states). A substance or object that emits nuclear radiation is said to be **radioactive**.

Two types of experimental evidence imply that Becquerel’s rays originate deep in the heart (or nucleus) of an atom. First, the radiation is found to be associated with certain elements, such as uranium. Radiation does not vary with chemical state—that is, uranium is radioactive whether it is in the form of an element or compound. In addition, radiation does not vary with temperature, pressure, or ionization state of the uranium atom. Since all of these factors affect electrons in an atom, the radiation cannot come from electron transitions, as atomic spectra do. The huge energy emitted during each event is the second piece of evidence that the radiation cannot be atomic. Nuclear radiation has energies of the order of 10^6 eV per event, which is much greater than the typical atomic energies (a few eV), such as that observed in spectra and chemical reactions, and more than ten times as high as the most energetic characteristic x rays. Becquerel did not vigorously pursue his discovery for very long. In 1898, Marie Curie (1867–

1934), then a graduate student married the already well-known French physicist Pierre Curie (1859–1906), began her doctoral study of Becquerel's rays. She and her husband soon discovered two new radioactive elements, which she named *polonium* (after her native land) and *radium* (because it radiates). These two new elements filled holes in the periodic table and, further, displayed much higher levels of radioactivity per gram of material than uranium. Over a period of four years, working under poor conditions and spending their own funds, the Curies processed more than a ton of uranium ore to isolate a gram of radium salt. Radium became highly sought after, because it was about two million times as radioactive as uranium. Curie's radium salt glowed visibly from the radiation that took its toll on them and other unaware researchers. Shortly after completing her Ph.D., both Curies and Becquerel shared the 1903 Nobel Prize in physics for their work on radioactivity. Pierre was killed in a horse cart accident in 1906, but Marie continued her study of radioactivity for nearly 30 more years. Awarded the 1911 Nobel Prize in chemistry for her discovery of two new elements, she remains the only person to win Nobel Prizes in physics and chemistry. Marie's radioactive fingerprints on some pages of her notebooks can still expose film, and she suffered from radiation-induced lesions. She died of leukemia likely caused by radiation, but she was active in research almost until her death in 1934. The following year, her daughter and son-in-law, Irene and Frederic Joliot-Curie, were awarded the Nobel Prize in chemistry for their discovery of artificially induced radiation, adding to a remarkable family legacy.

Alpha, Beta, and Gamma

Research begun by people such as New Zealander Ernest Rutherford soon after the discovery of nuclear radiation indicated that different types of rays are emitted. Eventually, three types were distinguished and named **alpha** (α), **beta**(β), and **gamma**(γ), because, like x-rays, their identities were initially unknown. [link] shows what happens if the rays are passed through a magnetic field. The γs are unaffected, while the α s and β s are deflected in opposite directions, indicating the α s are positive, the β s negative, and the γ s uncharged. Rutherford used both magnetic and electric fields to show that α s have a positive charge twice the magnitude of an electron, or $+2 \mid q_e \mid$. In the process, he found the α s charge to mass ratio to be several thousand times smaller than the electron’s. Later on, Rutherford collected α s from a radioactive source and passed an electric discharge through them, obtaining the spectrum of recently discovered helium gas. Among many important discoveries made by Rutherford and his collaborators was the proof that α *radiation is the emission of a helium nucleus*. Rutherford won the Nobel Prize in chemistry in 1908 for his early work. He continued to make important contributions until his death in 1934.



Alpha, beta, and gamma rays are passed through a magnetic field on the way to a phosphorescent screen. The α s and β s bend in opposite directions, while the γ s are unaffected, indicating a positive charge for α s, negative for β s, and neutral for γ s. Consistent results are obtained with electric fields. Collection of the radiation offers further confirmation from the direct measurement of excess charge.

Other researchers had already proved that β s are negative and have the same mass and same charge-to-mass ratio as the recently discovered electron. By 1902, it was recognized that *β radiation is the emission of an electron*. Although β s are electrons, they do not exist in the nucleus before it decays and are not ejected atomic electrons—the electron is created in the nucleus at the instant of decay.

Since γ s remain unaffected by electric and magnetic fields, it is natural to think they might be photons. Evidence for this grew, but it was not until 1914 that this was proved by Rutherford and collaborators. By scattering γ radiation from a crystal and observing interference, they demonstrated that *γ radiation is the emission of a high-energy photon by a nucleus*. In fact, γ radiation comes from the de-excitation of a nucleus, just as an x ray comes from the de-excitation of an atom. The names "γ ray" and "x ray" identify the source of the radiation. At the same energy, γ rays and x rays are otherwise identical.

Type of Radiation	**Range**
α -Particles	A sheet of paper, a few cm of air, fractions of a mm of tissue

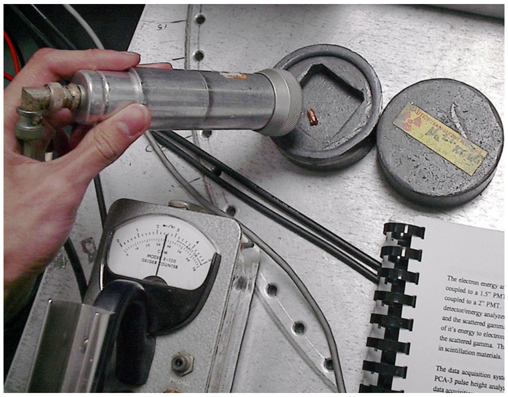
Type of Radiation	Range
β -Particles	A thin aluminum plate, or tens of cm of tissue
γ Rays	Several cm of lead or meters of concrete

Properties of Nuclear Radiation

Ionization and Range

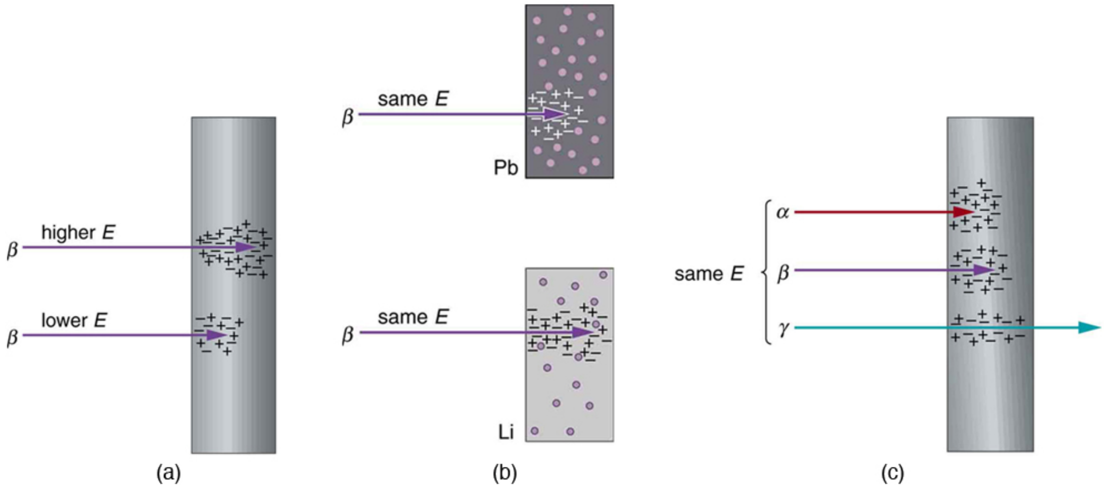
Two of the most important characteristics of α, β, and γ rays were recognized very early. All three types of nuclear radiation produce *ionization* in materials, but they penetrate different distances in materials—that is, they have different *ranges*. Let us examine why they have these characteristics and what are some of the consequences.

Like x rays, nuclear radiation in the form of α s, β s, and γ s has enough energy per event to ionize atoms and molecules in any material. The energy emitted in various nuclear decays ranges from a few keV to more than 10 MeV, while only a few eV are needed to produce ionization. The effects of x rays and nuclear radiation on biological tissues and other materials, such as solid state electronics, are directly related to the ionization they produce. All of them, for example, can damage electronics or kill cancer cells. In addition, methods for detecting x rays and nuclear radiation are based on ionization, directly or indirectly. All of them can ionize the air between the plates of a capacitor, for example, causing it to discharge. This is the basis of inexpensive personal radiation monitors, such as pictured in [link]. Apart from α, β, and γ, there are other forms of nuclear radiation as well, and these also produce ionization with similar effects. We define **ionizing radiation** as any form of radiation that produces ionization whether nuclear in origin or not, since the effects and detection of the radiation are related to ionization.



These dosimeters (literally, dose meters) are personal radiation monitors that detect the amount of radiation by the discharge of a rechargeable internal capacitor. The amount of discharge is related to the amount of ionizing radiation encountered, a measurement of dose. One dosimeter is shown in the charger. Its scale is read through an eyepiece on the top. (credit: L. Chang, Wikimedia Commons)

The **range of radiation** is defined to be the distance it can travel through a material. Range is related to several factors, including the energy of the radiation, the material encountered, and the type of radiation (see [link]). The higher the *energy*, the greater the range, all other factors being the same. This makes good sense, since radiation loses its energy in materials primarily by producing ionization in them, and each ionization of an atom or a molecule requires energy that is removed from the radiation. The amount of ionization is, thus, directly proportional to the energy of the particle of radiation, as is its range.



The penetration or range of radiation depends on its energy, the material it encounters, and the type of radiation. (a) Greater energy means greater range. (b) Radiation has a smaller range in materials with high electron density. (c) Alphas have the smallest range, betas have a greater range, and gammas penetrate the farthest.

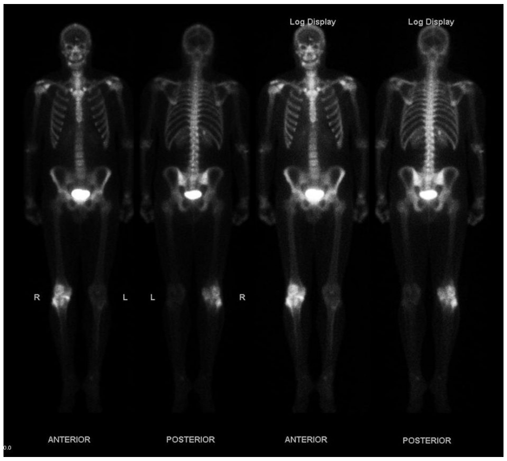
Radiation can be absorbed or shielded by materials, such as the lead aprons dentists drape on us when taking x rays. Lead is a particularly effective shield compared with other materials, such as plastic or air. How does the range of radiation depend on *material*? Ionizing radiation interacts best with charged particles in a material. Since electrons have small masses, they most readily absorb the energy of the radiation in collisions. The greater the density of a material and, in particular, the greater the density of electrons within a material, the smaller the range of radiation.

Note:
Collisions
Conservation of energy and momentum often results in energy transfer to a less massive object in a collision. This was discussed in detail in Work, Energy, and Energy Resources, for example.

Different *types* of radiation have different ranges when compared at the same energy and in the same material. Alphas have the shortest range, betas penetrate farther, and gammas have the greatest range. This is directly related to charge and speed of the particle or type of radiation. At a given energy, each α, β, or γ will produce the same number of ionizations in a material (each ionization requires a certain amount of energy on average). The more readily the particle produces ionization, the more quickly it will lose its energy. The effect of *charge* is as follows: The α has a charge of $+2q_e$, the β has a charge of $-q_e$, and the γ is uncharged. The electromagnetic force exerted by the α is thus twice as strong as that exerted by the β and it is more likely to produce ionization. Although chargeless, the γ does interact weakly because it is an electromagnetic wave, but it is less likely to produce ionization in any encounter. More quantitatively, the change in momentum Δp given to a particle in the material is $\Delta p = F\Delta t$, where F is the force the α, β, or γ exerts over a time Δt. The smaller the charge, the smaller is F and the smaller is the momentum (and energy) lost. Since the speed of alphas is about 5% to 10% of the speed of light, classical (non-relativistic) formulas apply.

The *speed* at which they travel is the other major factor affecting the range of α s, β s, and γ s. The faster they move, the less time they spend in the vicinity of an atom or a molecule, and the less likely they are to interact. Since α s and β s are particles with mass (helium nuclei and electrons, respectively), their energy is kinetic, given classically by $\frac{1}{2}mv^2$. The mass of the β particle is thousands of times less than that of the α s, so that β s must travel much faster than α s to have the same energy. Since β s move faster (most at relativistic speeds), they have less time to interact than α s. Gamma rays are photons, which must travel at the speed of light. They are even less likely to interact than a β, since they spend even less time near a given atom (and they have no charge). The range of γ s is thus greater than the range of β s.

Alpha radiation from radioactive sources has a range much less than a millimeter of biological tissues, usually not enough to even penetrate the dead layers of our skin. On the other hand, the same α radiation can penetrate a few centimeters of air, so mere distance from a source prevents α radiation from reaching us. This makes α radiation relatively safe for our body compared to β and γ radiation. Typical β radiation can penetrate a few millimeters of tissue or about a meter of air. Beta radiation is thus hazardous even when not ingested. The range of β s in lead is about a millimeter, and so it is easy to store β sources in lead radiation-proof containers. Gamma rays have a much greater range than either αs or βs. In fact, if a given thickness of material, like a lead brick, absorbs 90% of the γ s, then a second lead brick will only absorb 90% of what got through the first. Thus, γs do not have a well-defined range; we can only cut down the amount that gets through. Typically, γs can penetrate many meters of air, go right through our bodies, and are effectively shielded (that is, reduced in intensity to acceptable levels) by many centimeters of lead. One benefit of γ s is that they can be used as radioactive tracers (see [link]).



This image of the concentration of a radioactive tracer in a patient’s body reveals where the most active bone cells are, an indication of bone cancer. A short-lived radioactive substance that locates itself selectively is given to the patient, and the radiation is measured with an external detector. The emitted γ radiation has a sufficient range to leave the body—the range of α s and β s is too small for them to be observed outside the patient. (credit: Kieran Maher, Wikimedia Commons)

Note:
PhET Explorations: Beta Decay
Build an atom out of protons, neutrons, and electrons, and see how the element, charge, and mass change. Then play a game to test your ideas!

https://archive.cnx.org/specials/f0a27b96-f5c8-11e5-a22c-73f8c149bebf/beta-decay/#sim-multiple-atoms

Section Summary

- Some nuclei are radioactive—they spontaneously decay destroying some part of their mass and emitting energetic rays, a process called nuclear radioactivity.
- Nuclear radiation, like x rays, is ionizing radiation, because energy sufficient to ionize matter is emitted in each decay.
- The range (or distance traveled in a material) of ionizing radiation is directly related to the charge of the emitted particle and its energy, with greater-charge and lower-energy particles having the shortest ranges.
- Radiation detectors are based directly or indirectly upon the ionization created by radiation, as are the effects of radiation on living and inert materials.

Conceptual Questions

Exercise:

Problem:

Suppose the range for $5.0\text{ MeV}\alpha$ ray is known to be 2.0 mm in a certain material. Does this mean that every $5.0\text{ MeV}\alpha$ a ray that strikes this material travels 2.0 mm, or does the range have an average value with some statistical fluctuations in the distances traveled? Explain.

Exercise:

Problem:

What is the difference between γ rays and characteristic x rays? Is either necessarily more energetic than the other? Which can be the most energetic?

Exercise:

Problem:

Ionizing radiation interacts with matter by scattering from electrons and nuclei in the substance. Based on the law of conservation of momentum and energy, explain why electrons tend to absorb more energy than nuclei in these interactions.

Exercise:

Problem:

What characteristics of radioactivity show it to be nuclear in origin and not atomic?

Exercise:

Problem:

What is the source of the energy emitted in radioactive decay? Identify an earlier conservation law, and describe how it was modified to take such processes into account.

Exercise:

Problem:

Consider [link]. If an electric field is substituted for the magnetic field with positive charge instead of the north pole and negative charge instead of the south pole, in which directions will the α, β , and γ rays bend?

Exercise:

Problem:

Explain how an α particle can have a larger range in air than a β particle with the same energy in lead.

Exercise:

Problem:

Arrange the following according to their ability to act as radiation shields, with the best first and worst last. Explain your ordering in terms of how radiation loses its energy in matter.

(a) A solid material with low density composed of low-mass atoms.

(b) A gas composed of high-mass atoms.

(c) A gas composed of low-mass atoms.

(d) A solid with high density composed of high-mass atoms.

Exercise:

Problem:

Often, when people have to work around radioactive materials spills, we see them wearing white coveralls (usually a plastic material). What types of radiation (if any) do you think these suits protect the worker from, and how?

Glossary

alpha rays
: one of the types of rays emitted from the nucleus of an atom

beta rays
: one of the types of rays emitted from the nucleus of an atom

gamma rays

one of the types of rays emitted from the nucleus of an atom

ionizing radiation
radiation (whether nuclear in origin or not) that produces ionization whether nuclear in origin or not

nuclear radiation
rays that originate in the nuclei of atoms, the first examples of which were discovered by Becquerel

radioactivity
the emission of rays from the nuclei of atoms

radioactive
a substance or object that emits nuclear radiation

range of radiation
the distance that the radiation can travel through a material

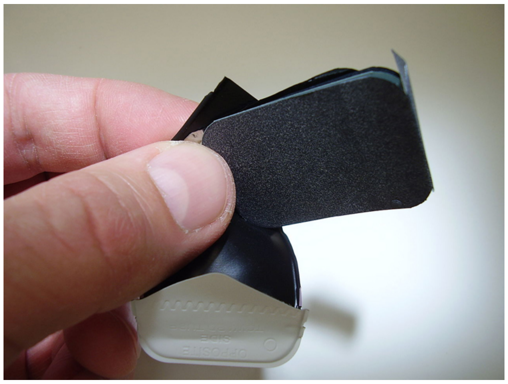
Radiation Detection and Detectors

- Explain the working principle of a Geiger tube.
- Define and discuss radiation detectors.

It is well known that ionizing radiation affects us but does not trigger nerve impulses. Newspapers carry stories about unsuspecting victims of radiation poisoning who fall ill with radiation sickness, such as burns and blood count changes, but who never felt the radiation directly. This makes the detection of radiation by instruments more than an important research tool. This section is a brief overview of radiation detection and some of its applications.

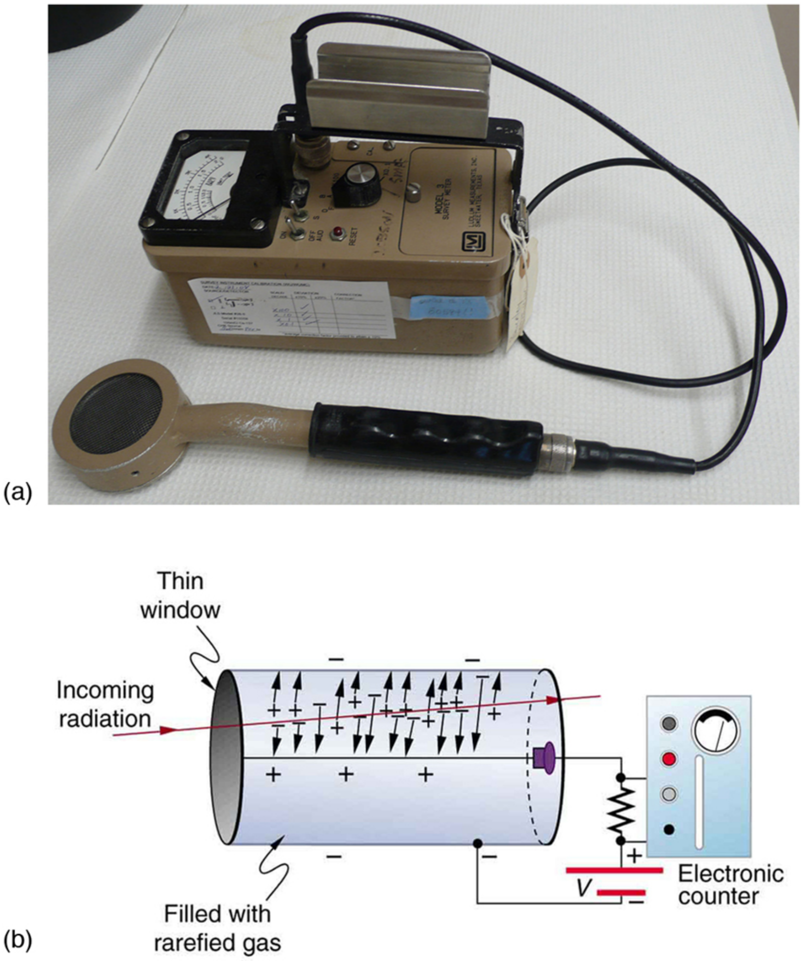
Human Application

The first direct detection of radiation was Becquerel's fogged photographic plate. Photographic film is still the most common detector of ionizing radiation, being used routinely in medical and dental x rays. Nuclear radiation is also captured on film, such as seen in [link]. The mechanism for film exposure by ionizing radiation is similar to that by photons. A quantum of energy interacts with the emulsion and alters it chemically, thus exposing the film. The quantum come from an α-particle, β-particle, or photon, provided it has more than the few eV of energy needed to induce the chemical change (as does all ionizing radiation). The process is not 100% efficient, since not all incident radiation interacts and not all interactions produce the chemical change. The amount of film darkening is related to exposure, but the darkening also depends on the type of radiation, so that absorbers and other devices must be used to obtain energy, charge, and particle-identification information.



Film badges contain film similar to that used in this dental x-ray film and is sandwiched between various absorbers to determine the penetrating ability of the radiation as well as the amount. (credit: Werneuchen, Wikimedia Commons)

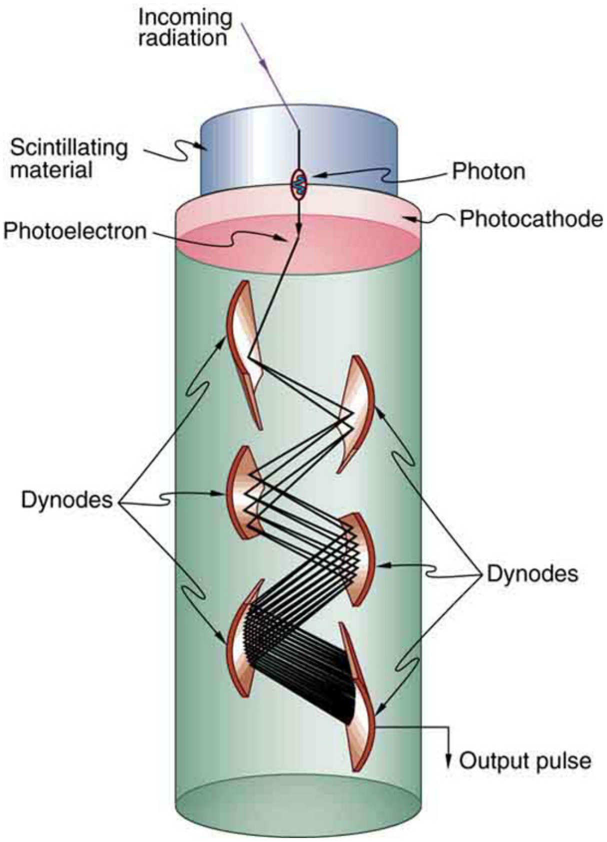
Another very common **radiation detector** is the **Geiger tube**. The clicking and buzzing sound we hear in dramatizations and documentaries, as well as in our own physics labs, is usually an audio output of events detected by a Geiger counter. These relatively inexpensive radiation detectors are based on the simple and sturdy Geiger tube, shown schematically in [link](b). A conducting cylinder with a wire along its axis is filled with an insulating gas so that a voltage applied between the cylinder and wire produces almost no current. Ionizing radiation passing through the tube produces free ion pairs that are attracted to the wire and cylinder, forming a current that is detected as a count. The word count implies that there is no information on energy, charge, or type of radiation with a simple Geiger counter. They do not detect every particle, since some radiation can pass through without producing enough ionization to be detected. However, Geiger counters are very useful in producing a prompt output that reveals the existence and relative intensity of ionizing radiation.



(a) Geiger counters such as this one are used for prompt monitoring of radiation levels, generally giving only relative intensity and not identifying the type or energy of the radiation. (credit: TimVickers, Wikimedia Commons) (b) Voltage applied between the cylinder and wire in a Geiger tube causes ions and electrons produced by radiation passing through the gas-filled cylinder to move towards them. The resulting current is detected and registered as a count.

Another radiation detection method records light produced when radiation interacts with materials. The energy of the radiation is sufficient to excite atoms in a material that may fluoresce, such as the phosphor used by Rutherford’s group. Materials called **scintillators** use a more complex collaborative process to convert radiation energy into light. Scintillators may be liquid or solid, and they can be very efficient. Their light output can provide information about the energy, charge, and type of radiation. Scintillator light flashes are very brief in duration, enabling the detection of a huge number of particles in short periods of time. Scintillator detectors are used in a variety of research and diagnostic applications. Among these are the detection by satellite-mounted equipment of the radiation from distant galaxies, the analysis of radiation from a person indicating body burdens, and the detection of exotic particles in accelerator laboratories.

Light from a scintillator is converted into electrical signals by devices such as the **photomultiplier** tube shown schematically in [link]. These tubes are based on the photoelectric effect, which is multiplied in stages into a cascade of electrons, hence the name photomultiplier. Light entering the photomultiplier strikes a metal plate, ejecting an electron that is attracted by a positive potential difference to the next plate, giving it enough energy to eject two or more electrons, and so on. The final output current can be made proportional to the energy of the light entering the tube, which is in turn proportional to the energy deposited in the scintillator. Very sophisticated information can be obtained with scintillators, including energy, charge, particle identification, direction of motion, and so on.



Photomultipliers use the photoelectric effect on the photocathode to convert the light output of a scintillator into an electrical signal. Each successive dynode has a more-positive potential than the last and attracts the ejected electrons, giving them more energy. The number of electrons is thus multiplied at each dynode, resulting in an easily detected output current.

Solid-state radiation detectors convert ionization produced in a semiconductor (like those found in computer chips) directly into an electrical signal. Semiconductors can be constructed that do not conduct current in one particular direction. When a voltage is applied in that direction, current flows only when ionization is produced by radiation, similar to what happens in a Geiger tube. Further, the amount of current in a solid-state detector is closely related to the energy deposited and, since the detector is solid, it can have a high efficiency (since ionizing radiation is stopped in a shorter distance in solids fewer particles escape detection). As with scintillators, very sophisticated information can be obtained from solid-state detectors.

Note:
PhET Explorations: Radioactive Dating Game
Learn about different types of radiometric dating, such as carbon dating. Understand how decay and half life work to enable radiometric dating to work. Play a game that tests your ability to match the percentage of the dating element that remains to the age of the object.

https://archive.cnx.org/specials/d709a8b0-068c-11e6-bcfb-f38266817c66/radioactive-dating-game/#sim-half-life

Section Summary

- Radiation detectors are based directly or indirectly upon the ionization created by radiation, as are the effects of radiation on living and inert materials.

Conceptual Questions

Exercise:

Problem:

Is it possible for light emitted by a scintillator to be too low in frequency to be used in a photomultiplier tube? Explain.

Problems & Exercises

Exercise:

Problem:

The energy of 30.0 eV is required to ionize a molecule of the gas inside a Geiger tube, thereby producing an ion pair. Suppose a particle of ionizing radiation deposits 0.500 MeV of energy in this Geiger tube. What maximum number of ion pairs can it create?

Solution:

1.67×10^4

Exercise:

Problem:

A particle of ionizing radiation creates 4000 ion pairs in the gas inside a Geiger tube as it passes through. What minimum energy was deposited, if 30.0 eV is required to create each ion pair?

Exercise:

Problem:

(a) Repeat [link], and convert the energy to joules or calories. (b) If all of this energy is converted to thermal energy in the gas, what is its temperature increase, assuming $50.0\ \text{cm}^3$ of ideal gas at 0.250-atm pressure? (The small answer is consistent with the fact that the energy is large on a quantum mechanical scale but small on a macroscopic scale.)

Exercise:

Problem:

Suppose a particle of ionizing radiation deposits 1.0 MeV in the gas of a Geiger tube, all of which goes to creating ion pairs. Each ion pair requires 30.0 eV of energy. (a) The applied voltage sweeps the ions out of the gas in 1.00 μs. What is the current? (b) This current is smaller than the actual current since the applied voltage in the Geiger tube accelerates the separated ions, which then create other ion pairs in subsequent collisions. What is the current if this last effect multiplies the number of ion pairs by 900?

Glossary

Geiger tube
: a very common radiation detector that usually gives an audio output

photomultiplier
: a device that converts light into electrical signals

radiation detector
: a device that is used to detect and track the radiation from a radioactive reaction

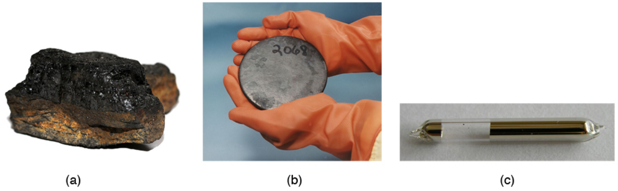
scintillators
: a radiation detection method that records light produced when radiation interacts with materials

solid-state radiation detectors
: semiconductors fabricated to directly convert incident radiation into electrical current

Substructure of the Nucleus

- Define and discuss the nucleus in an atom.
- Define atomic number.
- Define and discuss isotopes.
- Calculate the density of the nucleus.
- Explain nuclear force.

What is inside the nucleus? Why are some nuclei stable while others decay? (See [link].) Why are there different types of decay (α, β and γ)? Why are nuclear decay energies so large? Pursuing natural questions like these has led to far more fundamental discoveries than you might imagine.



Why is most of the carbon in this coal stable (a), while the uranium in the disk (b) slowly decays over billions of years? Why is cesium in this ampule (c) even less stable than the uranium, decaying in far less than 1/1,000,000 the time? What is the reason uranium and cesium undergo different types of decay (α and β, respectively)? (credits: (a) Bresson Thomas, Wikimedia Commons; (b) U.S. Department of Energy; (c) Tomihahndorf, Wikimedia Commons)

We have already identified **protons** as the particles that carry positive charge in the nuclei. However, there are actually *two* types of particles in the nuclei—the *proton* and the *neutron*, referred to collectively as **nucleons**, the constituents of nuclei. As its name implies, the **neutron** is a neutral particle ($q = 0$) that has nearly the same mass and intrinsic spin as the proton. [link] compares the masses of protons, neutrons, and electrons. Note how close the proton and neutron masses are, but the neutron is slightly more massive once you look past the third digit. Both nucleons are much more massive than an electron. In fact, $m_p = 1836 m_e$ (as noted in Medical Applications of Nuclear Physics and $m_n = 1839 m_e$.

[link] also gives masses in terms of mass units that are more convenient than kilograms on the atomic and nuclear scale. The first of these is the *unified* ***atomic mass*** *unit* (u), defined as
Equation:

$$1\text{ u} = 1.6605 \times 10^{-27}\text{ kg}.$$

This unit is defined so that a neutral carbon ^{12}C atom has a mass of exactly 12 u. Masses are also expressed in units of MeV/c^2. These units are very convenient when considering the conversion of mass into energy (and vice versa), as is so prominent in nuclear processes. Using $E = mc^2$ and units of m in MeV/c^2, we find that c^2 cancels and E comes out conveniently in MeV. For example, if the rest mass of a proton is converted entirely into energy, then
Equation:

$$E = mc^2 = (938.27\text{ MeV}/c^2)c^2 = 938.27\text{ MeV}.$$

It is useful to note that 1 u of mass converted to energy produces 931.5 MeV, or
Equation:

$$1\text{ u} = 931.5\text{ MeV}/c^2.$$

All properties of a nucleus are determined by the number of protons and neutrons it has. A specific combination of protons and neutrons is called a **nuclide** and is a unique nucleus. The following notation is used to represent a particular nuclide:
Equation:

$$^{A}_{Z}\text{X}_{N},$$

where the symbols A, X, Z , and N are defined as follows: The *number of protons in a nucleus* is the **atomic number** Z, as defined in Medical Applications of Nuclear Physics. X is the *symbol for the element*, such as Ca for calcium. However, once Z is known, the element is known; hence, Z and X are redundant. For example, $Z = 20$ is always calcium, and calcium always has $Z = 20$. N is the *number of neutrons* in a nucleus. In the notation for a nuclide, the subscript N is usually omitted. The symbol A is defined as the number of nucleons or the *total number of protons and neutrons,*

Equation:

$$A = N + Z,$$

where A is also called the **mass number**. This name for A is logical; the mass of an atom is nearly equal to the mass of its nucleus, since electrons have so little mass. The mass of the nucleus turns out to be nearly equal to the sum of the masses of the protons and neutrons in it, which is proportional to A. In this context, it is particularly convenient to express masses in units of u. Both protons and neutrons have masses close to 1 u, and so the mass of an atom is close to A u. For example, in an oxygen nucleus with eight protons and eight neutrons, $A = 16$, and its mass is 16 u. As noticed, the unified atomic mass unit is defined so that a neutral carbon atom (actually a ^{12}C atom) has a mass of *exactly* 12 u. Carbon was chosen as the standard, partly because of its importance in organic chemistry (see Appendix A).

Particle	Symbol	kg	u	MeVc^2
Proton	p	1.67262×10^{-27}	1.007276	938.27
Neutron	n	1.67493×10^{-27}	1.008665	939.57

Particle	**Symbol**	**kg**	**u**	**MeVc^2**
Electron	*e*	9.1094×10^{-31}	0.00054858	0.511

Masses of the Proton, Neutron, and Electron

Let us look at a few examples of nuclides expressed in the ${}^{A}_{Z}\text{X}_{N}$ notation. The nucleus of the simplest atom, hydrogen, is a single proton, or ${}^{1}_{1}\text{H}$ (the zero for no neutrons is often omitted). To check this symbol, refer to the periodic table —you see that the atomic number Z of hydrogen is 1. Since you are given that there are no neutrons, the mass number A is also 1. Suppose you are told that the helium nucleus or α particle has two protons and two neutrons. You can then see that it is written ${}^{4}_{2}\text{He}_{2}$. There is a scarce form of hydrogen found in nature called deuterium; its nucleus has one proton and one neutron and, hence, twice the mass of common hydrogen. The symbol for deuterium is, thus, ${}^{2}_{1}\text{H}_{1}$ (sometimes D is used, as for deuterated water D_2O). An even rarer—and radioactive—form of hydrogen is called tritium, since it has a single proton and two neutrons, and it is written ${}^{3}_{1}\text{H}_{2}$. These three varieties of hydrogen have nearly identical chemistries, but the nuclei differ greatly in mass, stability, and other characteristics. Nuclei (such as those of hydrogen) having the same Z and different N s are defined to be **isotopes** of the same element.

There is some redundancy in the symbols A, X, Z, and N . If the element X is known, then Z can be found in a periodic table and is always the same for a given element. If both A and X are known, then N can also be determined (first find Z; then, $N = A - Z$). Thus the simpler notation for nuclides is
Equation:

$${}^{A}\text{X},$$

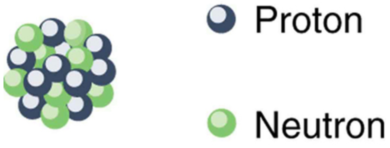
which is sufficient and is most commonly used. For example, in this simpler notation, the three isotopes of hydrogen are ${}^{1}\text{H}$, ${}^{2}\text{H}$, and ${}^{3}\text{H}$, while the α particle is ${}^{4}\text{He}$. We read this backward, saying helium-4 for ${}^{4}\text{He}$, or uranium-238 for ${}^{238}\text{U}$. So for ${}^{238}\text{U}$, should we need to know, we can determine that $Z = 92$ for uranium from the periodic table, and, thus, $N = 238 - 92 = 146$.

A variety of experiments indicate that a nucleus behaves something like a tightly packed ball of nucleons, as illustrated in [link]. These nucleons have large kinetic energies and, thus, move rapidly in very close contact. Nucleons can be separated by a large force, such as in a collision with another nucleus, but resist strongly being pushed closer together. The most compelling evidence that nucleons are closely packed in a nucleus is that the **radius of a nucleus**, r, is found to be given approximately by

Equation:

$$r = r_0 A^{1/3},$$

where $r_0 = 1.2\ \text{fm}$ and A is the mass number of the nucleus. Note that $r^3 \propto A$. Since many nuclei are spherical, and the volume of a sphere is $V = (4/3)\pi r^3$, we see that $V \propto A$ —that is, the volume of a nucleus is proportional to the number of nucleons in it. This is what would happen if you pack nucleons so closely that there is no empty space between them.



A model of the nucleus.

Nucleons are held together by nuclear forces and resist both being pulled apart and pushed inside one another. The volume of the nucleus is the sum of the volumes of the nucleons in it, here shown in different colors to represent protons and neutrons.

Example:
How Small and Dense Is a Nucleus?

(a) Find the radius of an iron-56 nucleus. (b) Find its approximate density in kg/m^3, approximating the mass of ^{56}Fe to be 56 u.

Strategy and Concept

(a) Finding the radius of ^{56}Fe is a straightforward application of $r = r_0 A^{1/3}$, given $A = 56$. (b) To find the approximate density, we assume the nucleus is spherical (this one actually is), calculate its volume using the radius found in part (a), and then find its density from $\rho = m/V$. Finally, we will need to convert density from units of u/fm^3 to kg/m^3.

Solution

(a) The radius of a nucleus is given by

Equation:

$$r = r_0 A^{1/3}.$$

Substituting the values for r_0 and A yields

Equation:

$$\begin{array}{rcl} r & = & (1.2\text{ fm})(56)^{1/3} = (1.2\text{ fm})(3.83) \\ & = & 4.6\text{ fm}. \end{array}$$

(b) Density is defined to be $\rho = m/V$, which for a sphere of radius r is

Equation:

$$\rho = \frac{m}{V} = \frac{m}{(4/3)\pi r^3}.$$

Substituting known values gives

Equation:

$$\begin{array}{rcl} \rho & = & \frac{56\text{ u}}{(1.33)(3.14)(4.6\text{ fm})^3} \\ & = & 0.138\text{ u/fm}^3. \end{array}$$

Converting to units of kg/m^3, we find

Equation:

$$\begin{array}{rcl} \rho & = & (0.138\text{ u/fm}^3)(1.66 \times 10^{-27}\text{ kg/u})\left(\frac{1\text{ fm}}{10^{-15}\text{ m}}\right) \\ & = & 2.3 \times 10^{17}\text{ kg/m}^3. \end{array}$$

Discussion

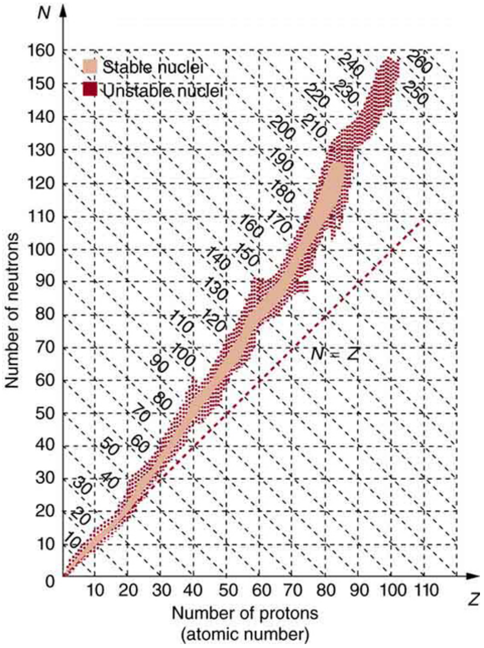
(a) The radius of this medium-sized nucleus is found to be approximately 4.6 fm, and so its diameter is about 10 fm, or 10^{-14} m. In our discussion of Rutherford's discovery of the nucleus, we noticed that it is about 10^{-15} m in diameter (which is for lighter nuclei), consistent with this result to an order of magnitude. The nucleus is much smaller in diameter than the typical atom, which has a diameter of the order of 10^{-10} m.
(b) The density found here is so large as to cause disbelief. It is consistent with earlier discussions we have had about the nucleus being very small and containing nearly all of the mass of the atom. Nuclear densities, such as found here, are about 2×10^{14} times greater than that of water, which has a density of "only" $10^3\ \text{kg/m}^3$. One cubic meter of nuclear matter, such as found in a neutron star, has the same mass as a cube of water 61 km on a side.

Nuclear Forces and Stability

What forces hold a nucleus together? The nucleus is very small and its protons, being positive, exert tremendous repulsive forces on one another. (The Coulomb force increases as charges get closer, since it is proportional to $1/r^2$, even at the tiny distances found in nuclei.) The answer is that two previously unknown forces hold the nucleus together and make it into a tightly packed ball of nucleons. These forces are called the *weak and strong nuclear forces*. Nuclear forces are so short ranged that they fall to zero strength when nucleons are separated by only a few fm. However, like glue, they are strongly attracted when the nucleons get close to one another. The strong nuclear force is about 100 times more attractive than the repulsive EM force, easily holding the nucleons together. Nuclear forces become extremely repulsive if the nucleons get too close, making nucleons strongly resist being pushed inside one another, something like ball bearings.

The fact that nuclear forces are very strong is responsible for the very large energies emitted in nuclear decay. During decay, the forces do work, and since work is force times the distance ($W = Fd$ cos θ), a large force can result in a large emitted energy. In fact, we know that there are *two* distinct nuclear forces because of the different types of nuclear decay—the strong nuclear force is responsible for α decay, while the weak nuclear force is responsible for β decay.

The many stable and unstable nuclei we have explored, and the hundreds we have not discussed, can be arranged in a table called the **chart of the nuclides**, a simplified version of which is shown in [link]. Nuclides are located on a plot of N versus Z. Examination of a detailed chart of the nuclides reveals patterns in the characteristics of nuclei, such as stability, abundance, and types of decay, analogous to but more complex than the systematics in the periodic table of the elements.



Simplified chart of the nuclides, a graph of N versus Z for known nuclides. The patterns of stable and unstable nuclides reveal characteristics of the nuclear forces. The dashed line is for $N = Z$. Numbers along diagonals are mass numbers A.

In principle, a nucleus can have any combination of protons and neutrons, but [link] shows a definite pattern for those that are stable. For low-mass nuclei, there is a strong tendency for N and Z to be nearly equal. This means that the nuclear force is more attractive when $N = Z$. More detailed examination reveals greater stability when N and Z are even numbers—nuclear forces are more attractive when neutrons and protons are in pairs. For increasingly higher masses, there are progressively more neutrons than protons in stable nuclei. This is due to the ever-growing repulsion between protons. Since nuclear forces are short ranged, and the Coulomb force is long ranged, an excess of neutrons keeps the protons a little farther apart, reducing Coulomb repulsion. Decay modes of nuclides out of the region of stability consistently produce nuclides closer to the region of stability. There are more stable nuclei having certain numbers of protons and neutrons, called **magic numbers**. Magic numbers indicate a shell structure for the nucleus in which closed shells are more stable. Nuclear shell theory has been very successful in explaining nuclear energy levels, nuclear decay, and the greater stability of nuclei with closed shells. We have been producing ever-heavier transuranic elements since the early 1940s, and we have now produced the element with $Z = 118$. There are theoretical predictions of an island of relative stability for nuclei with such high Z s.

The German-born American physicist Maria Goeppert Mayer (1906–1972) shared the 1963 Nobel Prize in physics with J. Jensen for the creation of the nuclear shell model. This successful nuclear model has nucleons filling shells analogous to electron shells in atoms. It was inspired by patterns observed in nuclear properties. (credit: Nobel Foundation via Wikimedia Commons)

Section Summary

- Two particles, both called nucleons, are found inside nuclei. The two types of nucleons are protons and neutrons; they are very similar, except that the proton is positively charged while the neutron is neutral. Some of their characteristics are given in [link] and compared with those of the electron. A mass unit convenient to atomic and nuclear processes is the unified atomic mass unit (u), defined to be
  Equation:

$$1\text{ u}=1.6605\times10^{-27}\text{ kg}=931.46\text{ MeV}/c^2.$$

- A nuclide is a specific combination of protons and neutrons, denoted by
  Equation:

$$^A_Z\mathrm{X}_N \text{ or simply} ^A\mathrm{X},$$

Z is the number of protons or atomic number, X is the symbol for the element, N is the number of neutrons, and A is the mass number or the total number of protons and neutrons,

Equation:

$$A = N + Z.$$

- Nuclides having the same Z but different N are isotopes of the same element.
- The radius of a nucleus, r, is approximately

  Equation:

  $$r = r_0 A^{1/3},$$

  where $r_0 = 1.2\ \mathrm{fm}$. Nuclear volumes are proportional to A. There are two nuclear forces, the weak and the strong. Systematics in nuclear stability seen on the chart of the nuclides indicate that there are shell closures in nuclei for values of Z and N equal to the magic numbers, which correspond to highly stable nuclei.

Conceptual Questions

Exercise:

Problem:

The weak and strong nuclear forces are basic to the structure of matter. Why we do not experience them directly?

Exercise:

Problem:

Define and make clear distinctions between the terms neutron, nucleon, nucleus, nuclide, and neutrino.

Exercise:

Problem:

What are isotopes? Why do different isotopes of the same element have similar chemistries?

Problems & Exercises

Exercise:

Problem:

Verify that a 2.3×10^{17} kg mass of water at normal density would make a cube 60 km on a side, as claimed in [link]. (This mass at nuclear density would make a cube 1.0 m on a side.)

Solution:
Equation:

$$\begin{aligned} m = \rho V = \rho d^3 \quad \Rightarrow \quad a &= \left(\frac{m}{\rho}\right)^{1/3} = \left(\frac{2.3 \times 10^{17}\ \text{kg}}{1000\ \text{kg/m}^3}\right)^{\frac{1}{3}} \\ &= 61 \times 10^3\ \text{m} = 61\ \text{km} \end{aligned}$$

Exercise:

Problem:

Find the length of a side of a cube having a mass of 1.0 kg and the density of nuclear matter, taking this to be $2.3 \times 10^{17}\ \text{kg/m}^3$.

Exercise:

Problem: What is the radius of an α particle?

Solution:

1.9 fm

Exercise:

Problem:

Find the radius of a ${}^{238}\text{Pu}$ nucleus. ${}^{238}\text{Pu}$ is a manufactured nuclide that is used as a power source on some space probes.

Exercise:

Problem:

(a) Calculate the radius of ${}^{58}\text{Ni}$, one of the most tightly bound stable nuclei.

(b) What is the ratio of the radius of ${}^{58}\text{Ni}$ to that of ${}^{258}\text{Ha}$, one of the largest nuclei ever made? Note that the radius of the largest nucleus is still much smaller than the size of an atom.

Solution:

(a) 4.6 fm

(b) $0.61\text{ to }1$

Exercise:

Problem:

The unified atomic mass unit is defined to be $1\text{ u}=1.6605\times10^{-27}\text{ kg}$. Verify that this amount of mass converted to energy yields 931.5 MeV. Note that you must use four-digit or better values for c and $\mid q_e \mid$.

Exercise:

Problem:

What is the ratio of the velocity of a β particle to that of an α particle, if they have the same nonrelativistic kinetic energy?

Solution:

$85.4\text{ to }1$

Exercise:

Problem:

If a 1.50-cm-thick piece of lead can absorb 90.0% of the γ rays from a radioactive source, how many centimeters of lead are needed to absorb all but 0.100% of the γ rays?

Exercise:

Problem:

The detail observable using a probe is limited by its wavelength. Calculate the energy of a γ-ray photon that has a wavelength of 1×10^{-16} m, small enough to detect details about one-tenth the size of a nucleon. Note that a photon having this energy is difficult to produce and interacts poorly with the nucleus, limiting the practicability of this probe.

Solution:

$12.4\ \text{GeV}$

Exercise:

Problem:

(a) Show that if you assume the average nucleus is spherical with a radius $r = r_0 A^{1/3}$, and with a mass of A u, then its density is independent of A.

(b) Calculate that density in u/fm^3 and kg/m^3, and compare your results with those found in [link] for ^{56}Fe.

Exercise:

Problem:

What is the ratio of the velocity of a 5.00-MeV β ray to that of an α particle with the same kinetic energy? This should confirm that βs travel much faster than αs even when relativity is taken into consideration. (See also [link].)

Solution:

19.3 to 1

Exercise:

Problem:

(a) What is the kinetic energy in MeV of a β ray that is traveling at $0.998c$? This gives some idea of how energetic a β ray must be to travel at nearly the same speed as a γ ray. (b) What is the velocity of the γ ray relative to the β ray?

Glossary

atomic mass
: the total mass of the protons, neutrons, and electrons in a single atom

atomic number
: number of protons in a nucleus

chart of the nuclides
: a table comprising stable and unstable nuclei

isotopes
: nuclei having the same Z and different Ns

magic numbers
: a number that indicates a shell structure for the nucleus in which closed shells are more stable

mass number
: number of nucleons in a nucleus

neutron
: a neutral particle that is found in a nucleus

nucleons
: the particles found inside nuclei

nucleus
: a region consisting of protons and neutrons at the center of an atom

nuclide

a type of atom whose nucleus has specific numbers of protons and neutrons

protons
the positively charged nucleons found in a nucleus

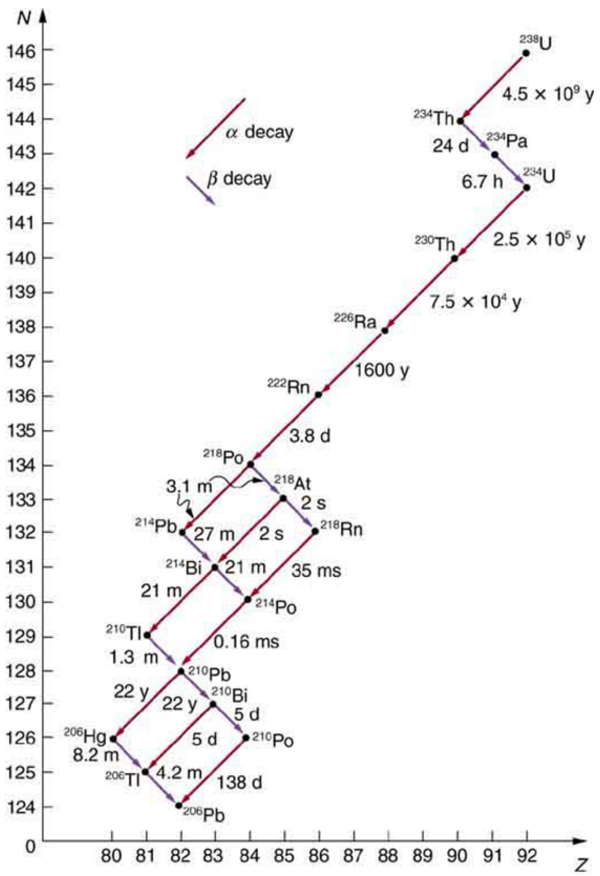
radius of a nucleus
the radius of a nucleus is $r = r_0 A^{1/3}$

Nuclear Decay and Conservation Laws

- Define and discuss nuclear decay.
- State the conservation laws.
- Explain parent and daughter nucleus.
- Calculate the energy emitted during nuclear decay.

Nuclear **decay** has provided an amazing window into the realm of the very small. Nuclear decay gave the first indication of the connection between mass and energy, and it revealed the existence of two of the four basic forces in nature. In this section, we explore the major modes of nuclear decay; and, like those who first explored them, we will discover evidence of previously unknown particles and conservation laws.

Some nuclides are stable, apparently living forever. Unstable nuclides decay (that is, they are radioactive), eventually producing a stable nuclide after many decays. We call the original nuclide the **parent** and its decay products the **daughters**. Some radioactive nuclides decay in a single step to a stable nucleus. For example, ^{60}Co is unstable and decays directly to ^{60}Ni, which is stable. Others, such as ^{238}U, decay to another unstable nuclide, resulting in a **decay series** in which each subsequent nuclide decays until a stable nuclide is finally produced. The decay series that starts from ^{238}U is of particular interest, since it produces the radioactive isotopes ^{226}Ra and ^{210}Po, which the Curies first discovered (see [link]). Radon gas is also produced (^{222}Rn in the series), an increasingly recognized naturally occurring hazard. Since radon is a noble gas, it emanates from materials, such as soil, containing even trace amounts of ^{238}U and can be inhaled. The decay of radon and its daughters produces internal damage. The ^{238}U decay series ends with ^{206}Pb, a stable isotope of lead.



The decay series produced by ^{238}U, the most common uranium isotope. Nuclides are graphed in the same manner as in the chart of nuclides. The type of decay for each member of the series is shown, as well as the half-lives. Note that some nuclides decay by more than one mode. You can see why radium and polonium are found in uranium ore. A stable isotope of lead is the end product of the series.

Note that the daughters of α decay shown in [link] always have two fewer protons and two fewer neutrons than the parent. This seems reasonable, since we know that α decay is the emission of a ^{4}He nucleus, which has two protons and two neutrons. The daughters of β decay have one less neutron and one more proton than their parent. Beta decay is a little more subtle, as we shall see. No γ decays are shown in the figure, because they do not produce a daughter that differs from the parent.

Alpha Decay

In **alpha decay**, a ^{4}He nucleus simply breaks away from the parent nucleus, leaving a daughter with two fewer protons and two fewer neutrons than the parent (see [link]). One example of α decay is shown in [link] for ^{238}U. Another nuclide that undergoes α decay is ^{239}Pu. The decay equations for these two nuclides are

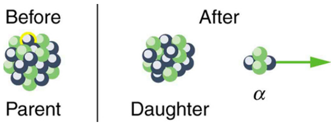
Equation:

$$^{238}\text{U} \rightarrow {}^{234}\text{Th}_{92}^{234} + {}^{4}\text{He}$$

and

Equation:

$$^{239}\text{Pu} \rightarrow {}^{235}\text{U} + {}^{4}\text{He}.$$



Alpha decay is the separation of a ^{4}He nucleus from the parent. The daughter nucleus has two fewer protons and two fewer neutrons than the parent. Alpha decay occurs spontaneously only if the daughter and ^{4}He nucleus have less total mass than the parent.

If you examine the periodic table of the elements, you will find that Th has $Z = 90$, two fewer than U, which has $Z = 92$. Similarly, in the second **decay equation**, we see that U has two fewer protons than Pu, which has $Z = 94$. The general rule for α decay is best written in the format ${}_{Z}^{A}\text{X}_{N}$. If a certain nuclide is known to α decay (generally this information must be looked up in a table of isotopes, such as in Appendix B), its α **decay equation** is

Equation:

$$^{A}_{Z}\text{X}_{N} \rightarrow {}^{A-4}_{Z-2}\text{Y}_{N-2} + {}^{4}_{2}\text{He}_{2} \quad (\alpha \text{ decay})$$

where Y is the nuclide that has two fewer protons than X, such as Th having two fewer than U. So if you were told that ^{239}Pu α decays and were asked to write the complete decay equation, you would first look up which element has two fewer protons (an atomic number two lower) and find that this is uranium. Then since four nucleons have broken away from the original 239, its atomic mass would be 235.

It is instructive to examine conservation laws related to α decay. You can see from the equation $^{A}_{Z}\text{X}_{N} \rightarrow {}^{A-4}_{Z-2}\text{Y}_{N-2} + {}^{4}_{2}\text{He}_{2}$ that total charge is conserved. Linear and angular momentum are conserved, too. Although conserved angular momentum is not of great consequence in this type of decay, conservation of linear momentum has interesting consequences. If the nucleus is at rest when it decays, its momentum is zero. In that case, the fragments must fly in opposite directions with equal-magnitude momenta so that total momentum remains zero. This results in the α particle carrying away most of the energy, as a bullet from a heavy rifle carries away most of the energy of the powder burned to shoot it. Total mass–energy is also conserved: the energy produced in the decay comes from conversion of a fraction of the original mass. As discussed in Atomic Physics, the general relationship is

Equation:

$$E = (\Delta m)c^2.$$

Here, E is the **nuclear reaction energy** (the reaction can be nuclear decay or any other reaction), and Δm is the difference in mass between initial and final products. When the final products have less total mass, Δm is positive, and the reaction releases energy (is exothermic). When the products have greater total mass, the reaction is endothermic (Δm is negative) and must be induced with an energy input. For α decay to be spontaneous, the decay products must have smaller mass than the parent.

Example:

Alpha Decay Energy Found from Nuclear Masses

Find the energy emitted in the α decay of ^{239}Pu.

Strategy

Nuclear reaction energy, such as released in α decay, can be found using the equation $E = (\Delta m)c^2$. We must first find Δm, the difference in mass between the parent nucleus and the products of the decay. This is easily done using masses given in Appendix A.

Solution

The decay equation was given earlier for ^{239}Pu ; it is

Equation:

$$^{239}\text{Pu} \rightarrow {}^{235}\text{U} + {}^{4}\text{He}.$$

Thus the pertinent masses are those of ^{239}Pu, ^{235}U, and the α particle or ^{4}He, all of which are listed in Appendix A. The initial mass was $m(^{239}\text{Pu}) = 239.052157\ \text{u}$. The final mass is the sum $m(^{235}\text{U})$+$m(^{4}\text{He})$= $235.043924\ \text{u} + 4.002602\ \text{u} = 239.046526\ \text{u}$. Thus,

Equation:

$$\begin{aligned} \Delta m &= m(^{239}\text{Pu}) - [m(^{235}\text{U}) + m(^{4}\text{He})] \\ &= 239.052157\ \text{u} - 239.046526\ \text{u} \\ &= 0.0005631\ \text{u}. \end{aligned}$$

Now we can find E by entering Δm into the equation:

Equation:

$$E = (\Delta m)c^2 = (0.005631\ \text{u})c^2.$$

We know $1\ \text{u} = 931.5\ \text{MeV}/c^2$, and so

Equation:

$$E = (0.005631)(931.5\ \text{MeV}/c^2)(c^2) = 5.25\ \text{MeV}.$$

Discussion

The energy released in this α decay is in the MeV range, about 10^6 times as great as typical chemical reaction energies, consistent with many previous discussions. Most of this energy becomes kinetic energy of the α particle (or ^{4}He nucleus), which moves away at high speed. The energy carried away by the recoil of the ^{235}U nucleus is much smaller in order to conserve momentum. The ^{235}U nucleus can be left in an excited state to later emit photons (γ rays). This decay is spontaneous and releases energy, because the products have less mass than the parent nucleus. The question of why the products have less mass will be discussed in Binding Energy. Note that the masses given in Appendix A are atomic masses of neutral atoms, including their electrons. The mass of the electrons is the same before and after α decay, and so their masses subtract out when finding Δm. In this case, there are 94 electrons before and after the decay.

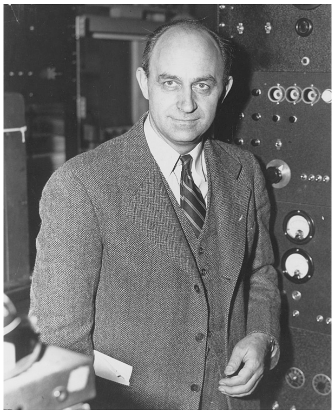
Beta Decay

There are actually *three* types of **beta decay**. The first discovered was “ordinary” beta decay and is called β^- decay or electron emission. The symbol β^- represents *an electron emitted in nuclear beta decay*. Cobalt-60 is a nuclide that β^- decays in the following manner:

Equation:

$$^{60}\text{Co} \rightarrow {}^{60}\text{Ni} + \beta^- + \text{neutrino}.$$

The **neutrino** is a particle emitted in beta decay that was unanticipated and is of fundamental importance. The neutrino was not even proposed in theory until more than 20 years after beta decay was known to involve electron emissions. Neutrinos are so difficult to detect that the first direct evidence of them was not obtained until 1953. Neutrinos are nearly massless, have no charge, and do not interact with nucleons via the strong nuclear force. Traveling approximately at the speed of light, they have little time to affect any nucleus they encounter. This is, owing to the fact that they have no charge (and they are not EM waves), they do not interact through the EM force. They do interact via the relatively weak and very short range weak nuclear force. Consequently, neutrinos escape almost any detector and penetrate almost any shielding. However, neutrinos do carry energy, angular momentum (they are fermions with half-integral spin), and linear momentum away from a beta decay. When accurate measurements of beta decay were made, it became apparent that energy, angular momentum, and linear momentum were not accounted for by the daughter nucleus and electron alone. Either a previously unsuspected particle was carrying them away, or three conservation laws were being violated. Wolfgang Pauli made a formal proposal for the existence of neutrinos in 1930. The Italian-born American physicist Enrico Fermi (1901–1954) gave neutrinos their name, meaning little neutral ones, when he developed a sophisticated theory of beta decay (see [link]). Part of Fermi’s theory was the identification of the weak nuclear force as being distinct from the strong nuclear force and in fact responsible for beta decay.



Enrico Fermi was nearly unique among 20th-century physicists—he made significant contributions both as an experimentalist and a theorist. His many contributions to theoretical physics included the identification of the weak nuclear force. The fermi (fm) is named after him, as are an entire class of subatomic particles (fermions), an element (Fermium), and a major research laboratory (Fermilab). His experimental work included studies of radioactivity, for which he won the 1938 Nobel Prize in physics, and creation of the first nuclear chain reaction. (credit: United States Department of Energy, Office of Public Affairs)

The neutrino also reveals a new conservation law. There are various families of particles, one of which is the electron family. We propose that the number of members of the electron family is constant in any process or any closed system. In our example of beta decay, there are no members of the electron family present before the decay, but after, there is an electron and a neutrino. So electrons are given an electron family number of $+1$. The neutrino in β^- decay is an **electron's antineutrino**, given the symbol ν_e, where ν is the Greek letter nu, and the subscript e means this neutrino is related to the electron. The bar indicates this is a particle of **antimatter**. (All particles have antimatter counterparts that are nearly identical except that they have the opposite charge. Antimatter is almost entirely absent on Earth, but it is found in nuclear decay and other nuclear and particle reactions as well as in outer space.) The electron's antineutrino ν_e, being antimatter, has an electron family number of –1. The total is zero, before and after the decay. The new conservation law, obeyed in all circumstances, states that the *total electron family number is constant*. An electron cannot be created without also creating an antimatter family member. This law is analogous to the conservation of charge in a situation where total charge is originally zero, and equal amounts of positive and negative charge must be created in a reaction to keep the total zero.

If a nuclide ${}_{Z}^{A}\text{X}_{N}$ is known to β^{-} decay, then its β^{-} decay equation is

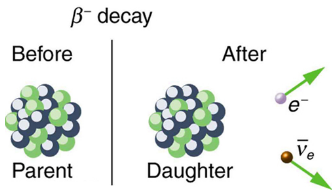
Equation:

$${}_{Z}^{A}\text{X}_{N} \rightarrow {}_{Z+1}^{A}\text{Y}_{N-1} + \beta^{-} + \bar{\nu}_{e} \;(\beta^{-} \text{ decay}),$$

where Y is the nuclide having one more proton than X (see [link]). So if you know that a certain nuclide β^{-} decays, you can find the daughter nucleus by first looking up Z for the parent and then determining which element has atomic number $Z+1$. In the example of the β^{-} decay of ^{60}Co given earlier, we see that $Z=27$ for Co and $Z=28$ is Ni. It is as if one of the neutrons in the parent nucleus decays into a proton, electron, and neutrino. In fact, neutrons outside of nuclei do just that—they live only an average of a few minutes and β^{-} decay in the following manner:

Equation:

$$\text{n} \rightarrow \text{p} + \beta^{-} + \bar{\nu}_{e}.$$



In β^{-} decay, the parent nucleus emits an electron and an antineutrino. The daughter nucleus has one more proton and one less neutron than its parent. Neutrinos interact so weakly that they are almost never directly observed, but they play a fundamental role in particle physics.

We see that charge is conserved in β^- decay, since the total charge is Z before and after the decay. For example, in ^{60}Co decay, total charge is 27 before decay, since cobalt has $Z = 27$. After decay, the daughter nucleus is Ni, which has $Z = 28$, and there is an electron, so that the total charge is also $28 + (-1)$ or 27. Angular momentum is conserved, but not obviously (you have to examine the spins and angular momenta of the final products in detail to verify this). Linear momentum is also conserved, again imparting most of the decay energy to the electron and the antineutrino, since they are of low and zero mass, respectively. Another new conservation law is obeyed here and elsewhere in nature. *The total number of nucleons A is conserved*. In ^{60}Co decay, for example, there are 60 nucleons before and after the decay. Note that total A is also conserved in α decay. Also note that the total number of protons changes, as does the total number of neutrons, so that total Z and total N are *not* conserved in β^- decay, as they are in α decay. Energy released in β^- decay can be calculated given the masses of the parent and products.

Example:

β^- Decay Energy from Masses

Find the energy emitted in the β^- decay of ^{60}Co.

Strategy and Concept

As in the preceding example, we must first find Δm, the difference in mass between the parent nucleus and the products of the decay, using masses given in Appendix A. Then the emitted energy is calculated as before, using $E = (\Delta m)c^2$. The initial mass is just that of the parent nucleus, and the final mass is that of the daughter nucleus and the electron created in the decay. The neutrino is massless, or nearly so. However, since the masses given in Appendix A are for neutral atoms, the daughter nucleus has one more electron than the parent, and so the extra electron mass that corresponds to the β^- is included in the atomic mass of Ni. Thus,

Equation:

$$\Delta m = m(^{60}\text{Co}) - m(^{60}\text{Ni}\ ).$$

Solution

The β^- decay equation for ^{60}Co is

Equation:

$$^{60}_{27}\text{Co}_{33} \rightarrow {}^{60}_{28}\text{Ni}_{32} + \beta^- + \nu_e.$$

As noticed,

Equation:

$$\Delta m = m(^{60}\text{Co}) - m(^{60}\text{Ni}\ ).$$

Entering the masses found in Appendix A gives

Equation:

$$\Delta m = 59.933820\text{ u} - 59.930789\text{ u} = 0.003031\text{ u}.$$

Thus,

Equation:

$$E = (\Delta m)c^2 = (0.003031\text{ u})c^2.$$

Using $1\text{ u} = 931.5\text{ MeV}/c^2$, we obtain

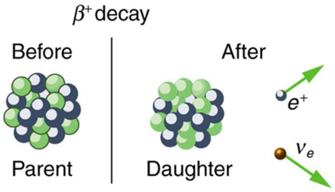
Equation:

$$E = (0.003031)(931.5\text{ MeV}/c^2)(c^2) = 2.82\text{ MeV}.$$

Discussion and Implications

Perhaps the most difficult thing about this example is convincing yourself that the β^- mass is included in the atomic mass of ^{60}Ni. Beyond that are other implications. Again the decay energy is in the MeV range. This energy is shared by all of the products of the decay. In many ^{60}Co decays, the daughter nucleus ^{60}Ni is left in an excited state and emits photons (γ rays). Most of the remaining energy goes to the electron and neutrino, since the recoil kinetic energy of the daughter nucleus is small. One final note: the electron emitted in β^- decay is created in the nucleus at the time of decay.

The second type of beta decay is less common than the first. It is β^+decay. Certain nuclides decay by the emission of a *positive* electron. This is **antielectron** or **positron decay** (see [link]).



β^+ decay is the emission of a positron that eventually finds an electron to annihilate, characteristically producing gammas in opposite directions.

The antielectron is often represented by the symbol e^+, but in beta decay it is written as β^+ to indicate the antielectron was emitted in a nuclear decay. Antielectrons are the antimatter counterpart to electrons, being nearly identical, having the same mass, spin, and so on, but having a positive charge and an electron family number of –1. When a **positron** encounters an electron, there is a mutual annihilation in which all the mass of the antielectron-electron pair is converted into pure photon energy. (The reaction, $e^+ + e^- \rightarrow \gamma + \gamma$, conserves electron family number as well as all other conserved quantities.) If a nuclide ${}_{Z}^{A}\text{X}_{N}$ is known to β^+ decay, then its β^+**decay equation** is

Equation:

$${}_{Z}^{A}\text{X}_{N} \rightarrow {}_{Z-1}^{A}\text{Y}_{N+1} + \beta^+ + \nu_e \text{ (}\beta^+\text{ decay),}$$

where Y is the nuclide having one less proton than X (to conserve charge) and ν_e is the symbol for the **electron's neutrino**, which has an electron family number of $+1$. Since an antimatter member of the electron family (the β^+) is created in the decay, a matter member of the family (here the ν_e) must also be created. Given, for example, that ${}^{22}\text{Na}$ β^+ decays, you can write its full decay equation by first finding that $Z = 11$ for ${}^{22}\text{Na}$, so that the daughter nuclide will have $Z = 10$, the atomic number for neon. Thus the β^+ decay equation for ${}^{22}\text{Na}$ is

Equation:

$${}_{11}^{22}\text{Na}_{11} \rightarrow {}_{10}^{22}\text{Ne}_{12} + \beta^+ + \nu_e.$$

In β^+ decay, it is as if one of the protons in the parent nucleus decays into a neutron, a positron, and a neutrino. Protons do not do this outside of the nucleus, and so the decay is due to the complexities of the nuclear force. Note again that the total number of nucleons is constant in this and any other reaction. To find the energy emitted in β^+ decay, you must again count the number of electrons in the neutral atoms, since atomic masses are used. The daughter has one less electron than the parent, and one electron mass is created in the decay. Thus, in β^+ decay,

Equation:

$$\Delta m = m(\text{parent}) - [m(\text{daughter}) + 2m_e],$$

since we use the masses of neutral atoms.

Electron capture is the third type of beta decay. Here, a nucleus captures an inner-shell electron and undergoes a nuclear reaction that has the same effect as β^+ decay. Electron capture is sometimes denoted by the letters EC. We know that electrons cannot reside in the nucleus, but this is a nuclear reaction that consumes the electron and occurs spontaneously only when the products have less mass than the parent plus the electron. If a nuclide ${}_{Z}^{A}\text{X}_{N}$ is known to undergo electron capture, then its **electron capture equation** is

Equation:

$$ {}_{Z}^{A}\text{X}_{N} + e^{-} \rightarrow {}_{Z-1}^{A}\text{Y}_{N+1} + \nu_{e} (\text{electron capture, or EC}). $$

Any nuclide that can β^{+} decay can also undergo electron capture (and often does both). The same conservation laws are obeyed for EC as for β^{+} decay. It is good practice to confirm these for yourself.

All forms of beta decay occur because the parent nuclide is unstable and lies outside the region of stability in the chart of nuclides. Those nuclides that have relatively more neutrons than those in the region of stability will β^{-} decay to produce a daughter with fewer neutrons, producing a daughter nearer the region of stability. Similarly, those nuclides having relatively more protons than those in the region of stability will β^{-} decay or undergo electron capture to produce a daughter with fewer protons, nearer the region of stability.

Gamma Decay

Gamma decay is the simplest form of nuclear decay—it is the emission of energetic photons by nuclei left in an excited state by some earlier process. Protons and neutrons in an excited nucleus are in higher orbitals, and they fall to lower levels by photon emission (analogous to electrons in excited atoms). Nuclear excited states have lifetimes typically of only about 10^{-14} s, an indication of the great strength of the forces pulling the nucleons to lower states. The γ decay equation is simply
Equation:

$$ {}_{Z}^{A}\text{X}_{N}^{*} \rightarrow {}_{Z}^{A}\text{X}_{N} + \gamma_{1} + \gamma_{2} + \cdots \ (\gamma \text{ decay}) $$

where the asterisk indicates the nucleus is in an excited state. There may be one or more γ s emitted, depending on how the nuclide de-excites. In radioactive decay, γ emission is common and is preceded by γ or β decay. For example, when ^{60}Co β^{-} decays, it most often leaves the daughter nucleus in an excited state, written ^{60}Ni*. Then the nickel nucleus quickly γ decays by the emission of two penetrating γ s:
Equation:

$$ ^{60}\text{Ni}^{*} \rightarrow {}^{60}\text{Ni} + \gamma_{1} + \gamma_{2}. $$

These are called cobalt γ rays, although they come from nickel—they are used for cancer therapy, for example. It is again constructive to verify the conservation laws for gamma decay. Finally, since γ decay does not change the nuclide to another species, it is not prominently featured in charts of decay series, such as that in [link].

There are other types of nuclear decay, but they occur less commonly than α, β, and γ decay. Spontaneous fission is the most important of the other forms of nuclear decay because of its applications in nuclear power and weapons. It is covered in the next chapter.

Section Summary

- When a parent nucleus decays, it produces a daughter nucleus following rules and conservation laws. There are three major types of nuclear decay, called alpha $(\alpha),$ beta $(\beta),$ and gamma (γ). The α decay equation is
  Equation:
  $${}_{Z}^{A}\mathrm{X}_{N} \to {}_{Z-2}^{A-4}\mathrm{Y}_{N-2} + {}_{2}^{4}\mathrm{He}_{2}.$$
- Nuclear decay releases an amount of energy E related to the mass destroyed Δm by
  Equation:
  $$E = (\Delta m)c^2.$$
- There are three forms of beta decay. The β^-decay equation is
  Equation:
  $${}_{Z}^{A}\mathrm{X}_{N} \to {}_{Z+1}^{A}\mathrm{Y}_{N-1} + \beta^- + \nu_e.$$
- The β^+ decay equation is
  Equation:
  $${}_{Z}^{A}\mathrm{X}_{N} \to {}_{Z-1}^{A}\mathrm{Y}_{N+1} + \beta^+ + \nu_e.$$
- The electron capture equation is
  Equation:
  $${}_{Z}^{A}\mathrm{X}_{N} + e^- \to {}_{Z-1}^{A}\mathrm{Y}_{N+1} + \nu_e.$$
- β^- is an electron, β^+ is an antielectron or positron, ν_e represents an electron's neutrino, and ν_e is an electron's antineutrino. In addition to all previously known conservation laws, two new ones arise— conservation of electron family number and conservation of the total number of nucleons. The γ decay equation is
  Equation:
  $${}_{Z}^{A}\mathrm{X}_{N}^{*} \to {}_{Z}^{A}\mathrm{X}_{N} + \gamma_1 + \gamma_2 + \cdots$$

  γ is a high-energy photon originating in a nucleus.

Conceptual Questions

Exercise:

Problem:

Star Trek fans have often heard the term “antimatter drive.” Describe how you could use a magnetic field to trap antimatter, such as produced by nuclear decay, and later combine it with matter to produce energy. Be specific about the type of antimatter, the need for vacuum storage, and the fraction of matter converted into energy.

Exercise:

Problem:

What conservation law requires an electron’s neutrino to be produced in electron capture? Note that the electron no longer exists after it is captured by the nucleus.

Exercise:

Problem:

Neutrinos are experimentally determined to have an extremely small mass. Huge numbers of neutrinos are created in a supernova at the same time as massive amounts of light are first produced. When the 1987A supernova occurred in the Large Magellanic Cloud, visible primarily in the Southern Hemisphere and some 100,000 light-years away from Earth, neutrinos from the explosion were observed at about the same time as the light from the blast. How could the relative arrival times of neutrinos and light be used to place limits on the mass of neutrinos?

Exercise:

Problem:

What do the three types of beta decay have in common that is distinctly different from alpha decay?

Problems & Exercises

In the following eight problems, write the complete decay equation for the given nuclide in the complete ${}_{Z}^{A}\text{X}_{N}$ notation. Refer to the periodic table for values of Z.

Exercise:

Problem:

β^- decay of ${}^{3}\text{H}$ (tritium), a manufactured isotope of hydrogen used in some digital watch displays, and manufactured primarily for use in hydrogen bombs.

Solution:

Equation:

$${}_{1}^{3}\text{H}_{2} \rightarrow {}_{2}^{3}\text{He}_{1} + \beta^- + \nu_e$$

Exercise:

Problem:

β^- decay of ^{40}K, a naturally occurring rare isotope of potassium responsible for some of our exposure to background radiation.

Exercise:

Problem: β^+ decay of ^{50}Mn.

Solution:
Equation:

$$^{50}_{25}M_{25} \rightarrow {}^{50}_{24}\text{Cr}_{26} + \beta^+ + \nu_e$$

Exercise:

Problem: β^+ decay of ^{52}Fe.

Exercise:

Problem: Electron capture by ^{7}Be.

Solution:
Equation:

$$^{7}_{4}\text{Be}_3 + e^- \rightarrow {}^{7}_{3}\text{Li}_4 + \nu_e$$

Exercise:

Problem: Electron capture by ^{106}In.

Exercise:

Problem:

α decay of ^{210}Po, the isotope of polonium in the decay series of ^{238}U that was discovered by the Curies. A favorite isotope in physics labs, since it has a short half-life and decays to a stable nuclide.

Solution:
Equation:

$$^{210}_{84}\text{Po}_{126} \rightarrow {}^{206}_{82}\text{Pb}_{124} + {}^{4}_{2}\text{He}_2$$

Exercise:

Problem:

α decay of $^{226}\mathrm{Ra}$, another isotope in the decay series of $^{238}\mathrm{U}$, first recognized as a new element by the Curies. Poses special problems because its daughter is a radioactive noble gas.

In the following four problems, identify the parent nuclide and write the complete decay equation in the $^{A}_{Z}\mathrm{X}_{N}$ notation. Refer to the periodic table for values of Z.

Exercise:

Problem:

β^{-} decay producing $^{137}\mathrm{Ba}$. The parent nuclide is a major waste product of reactors and has chemistry similar to potassium and sodium, resulting in its concentration in your cells if ingested.

Solution:

Equation:

$$^{137}_{55}\mathrm{Cs}_{82} \rightarrow {}^{137}_{56}\mathrm{Ba}_{81} + \beta^{-} + \nu_{e}$$

Exercise:

Problem:

β^{-} decay producing $^{90}\mathrm{Y}$. The parent nuclide is a major waste product of reactors and has chemistry similar to calcium, so that it is concentrated in bones if ingested ($^{90}\mathrm{Y}$ is also radioactive.)

Exercise:

Problem:

α decay producing $^{228}\mathrm{Ra}$. The parent nuclide is nearly 100% of the natural element and is found in gas lantern mantles and in metal alloys used in jets ($^{228}\mathrm{Ra}$ is also radioactive).

Solution:

Equation:

$$^{232}_{90}\mathrm{Th}_{142} \rightarrow {}^{228}_{88}\mathrm{Ra}_{140} + {}^{4}_{2}\mathrm{He}_{2}$$

Exercise:

Problem:

α decay producing ^{208}Pb. The parent nuclide is in the decay series produced by ^{232}Th, the only naturally occurring isotope of thorium.

Exercise:

Problem:

When an electron and positron annihilate, both their masses are destroyed, creating two equal energy photons to preserve momentum. (a) Confirm that the annihilation equation $e^{+} + e^{-} \rightarrow \gamma + \gamma$ conserves charge, electron family number, and total number of nucleons. To do this, identify the values of each before and after the annihilation. (b) Find the energy of each γ ray, assuming the electron and positron are initially nearly at rest. (c) Explain why the two γ rays travel in exactly opposite directions if the center of mass of the electron-positron system is initially at rest.

Solution:

(a)
$\text{charge:}(+1) + (-1) = 0; \quad \text{electron family number: } (+1) + (-1) = 0; \quad A\text{: } 0 + 0 = 0$

(b) 0.511 MeV

(c) The two γ rays must travel in exactly opposite directions in order to conserve momentum, since initially there is zero momentum if the center of mass is initially at rest.

Exercise:

Problem:

Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for α decay given in the equation ${}_{Z}^{A}\text{X}_{N} \rightarrow {}_{Z-2}^{A-4}\text{Y}_{N-2} + {}_{2}^{4}\text{He}_{2}$. To do this, identify the values of each before and after the decay.

Exercise:

Problem:

Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for β^{-} decay given in the equation ${}_{Z}^{A}\text{X}_{N} \rightarrow {}_{Z+1}^{A}\text{Y}_{N-1} + \beta^{-} + \nu_{e}$. To do this, identify the values of each before and after the decay.

Solution:
Equation:

$$Z = (Z + 1) - 1; \quad A = A; \quad \text{efn} : 0 = (+1) + (-1)$$

Exercise:

Problem:

Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for β^- decay given in the equation ${}_Z^A\text{X}_N \to {}_{Z-1}^{A}\text{Y}_{N-1} + \beta^- + \nu_e$. To do this, identify the values of each before and after the decay.

Exercise:

Problem:

Confirm that charge, electron family number, and the total number of nucleons are all conserved by the rule for electron capture given in the equation ${}_Z^A\text{X}_N + e^- \to {}_{Z-1}^{A}\text{Y}_{N+1} + \nu_e$. To do this, identify the values of each before and after the capture.

Solution:
Equation:

$$Z-1=Z-1;\quad A=A;\quad \text{efn}:(+1)=(+1)$$

Exercise:

Problem:

A rare decay mode has been observed in which ${}^{222}\text{Ra}$ emits a ${}^{14}\text{C}$ nucleus. (a) The decay equation is ${}^{222}\text{Ra} \to^{A}\text{X}+{}^{14}\text{C}$. Identify the nuclide ${}^{A}\text{X}$. (b) Find the energy emitted in the decay. The mass of ${}^{222}\text{Ra}$ is 222.015353 u.

Exercise:

Problem: (a) Write the complete α decay equation for ${}^{226}\text{Ra}$.

(b) Find the energy released in the decay.

Solution:

(a) ${}_{88}^{226}\text{Ra}_{138} \to {}_{86}^{222}\text{Rn}_{136} + {}_{2}^{4}\text{He}_2$

(b) 4.87 MeV

Exercise:

Problem: (a) Write the complete α decay equation for ${}^{249}\text{Cf}$.

(b) Find the energy released in the decay.

Exercise:

Problem:

(a) Write the complete β^- decay equation for the neutron. (b) Find the energy released in the decay.

Solution:

(a) $\mathrm{n} \rightarrow \mathrm{p} + \beta^- + \nu_e$

(b)) 0.783 MeV

Exercise:

Problem:

(a) Write the complete β^- decay equation for $^{90}\mathrm{Sr}$, a major waste product of nuclear reactors. (b) Find the energy released in the decay.

Exercise:

Problem:

Calculate the energy released in the β^+ decay of $^{22}\mathrm{Na}$, the equation for which is given in the text. The masses of $^{22}\mathrm{Na}$ and $^{22}\mathrm{Ne}$ are 21.994434 and 21.991383 u, respectively.

Solution:

1.82 MeV

Exercise:

Problem: (a) Write the complete β^+ decay equation for $^{11}\mathrm{C}$.

(b) Calculate the energy released in the decay. The masses of $^{11}\mathrm{C}$ and $^{11}\mathrm{B}$ are 11.011433 and 11.009305 u, respectively.

Exercise:

Problem: (a) Calculate the energy released in the α decay of $^{238}\mathrm{U}$.

(b) What fraction of the mass of a single $^{238}\mathrm{U}$ is destroyed in the decay? The mass of $^{234}\mathrm{Th}$ is 234.043593 u.

(c) Although the fractional mass loss is large for a single nucleus, it is difficult to observe for an entire macroscopic sample of uranium. Why is this?

Solution:

(a) $4.274\ \mathrm{MeV}$

(b) 1.927×10^{-5}

(c) Since U-238 is a slowly decaying substance, only a very small number of nuclei decay on human timescales; therefore, although those nuclei that decay lose a noticeable fraction of their mass, the change in the total mass of the sample is not detectable for a macroscopic sample.

Exercise:

Problem: (a) Write the complete reaction equation for electron capture by $^{7}\mathrm{Be}$.

(b) Calculate the energy released.

Exercise:

Problem: (a) Write the complete reaction equation for electron capture by $^{15}\mathrm{O}$.

(b) Calculate the energy released.

Solution:

(a) ${}^{15}_{8}\mathrm{O}_7 + e^- \rightarrow {}^{15}_{7}\mathrm{N}_8 + \nu_e$

(b) 2.754 MeV

Glossary

parent
: the original state of nucleus before decay

daughter
: the nucleus obtained when parent nucleus decays and produces another nucleus following the rules and the conservation laws

positron
: the particle that results from positive beta decay; also known as an antielectron

decay
: the process by which an atomic nucleus of an unstable atom loses mass and energy by emitting ionizing particles

alpha decay
: type of radioactive decay in which an atomic nucleus emits an alpha particle

beta decay
type of radioactive decay in which an atomic nucleus emits a beta particle

gamma decay
type of radioactive decay in which an atomic nucleus emits a gamma particle

decay equation
the equation to find out how much of a radioactive material is left after a given period of time

nuclear reaction energy
the energy created in a nuclear reaction

neutrino
an electrically neutral, weakly interacting elementary subatomic particle

electron’s antineutrino
antiparticle of electron’s neutrino

positron decay
type of beta decay in which a proton is converted to a neutron, releasing a positron and a neutrino

antielectron
another term for positron

decay series
process whereby subsequent nuclides decay until a stable nuclide is produced

electron’s neutrino
a subatomic elementary particle which has no net electric charge

antimatter
composed of antiparticles

electron capture
the process in which a proton-rich nuclide absorbs an inner atomic electron and simultaneously emits a neutrino

electron capture equation
equation representing the electron capture

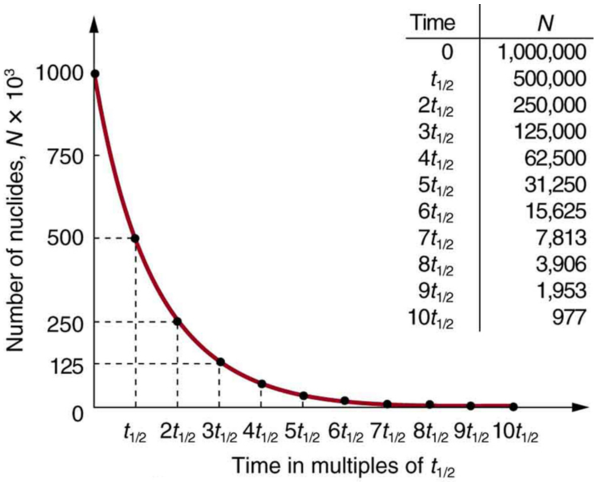
Half-Life and Activity

- Define half-life.
- Define dating.
- Calculate age of old objects by radioactive dating.

Unstable nuclei decay. However, some nuclides decay faster than others. For example, radium and polonium, discovered by the Curies, decay faster than uranium. This means they have shorter lifetimes, producing a greater rate of decay. In this section we explore half-life and activity, the quantitative terms for lifetime and rate of decay.

Half-Life

Why use a term like half-life rather than lifetime? The answer can be found by examining [link], which shows how the number of radioactive nuclei in a sample decreases with time. The *time in which half of the original number of nuclei decay* is defined as the **half-life**, $t_{1/2}$. Half of the remaining nuclei decay in the next half-life. Further, half of that amount decays in the following half-life. Therefore, the number of radioactive nuclei decreases from N to $N/2$ in one half-life, then to $N/4$ in the next, and to $N/8$ in the next, and so on. If N is a large number, then *many* half-lives (not just two) pass before all of the nuclei decay. Nuclear decay is an example of a purely statistical process. A more precise definition of half-life is that *each nucleus has a 50% chance of living for a time equal to one half-life* $t_{1/2}$. Thus, if N is reasonably large, half of the original nuclei decay in a time of one half-life. If an individual nucleus makes it through that time, it still has a 50% chance of surviving through another half-life. Even if it happens to make it through hundreds of half-lives, it still has a 50% chance of surviving through one more. The probability of decay is the same no matter when you start counting. This is like random coin flipping. The chance of heads is 50%, no matter what has happened before.



Time	N
0	1,000,000
$t_{1/2}$	500,000
$2t_{1/2}$	250,000
$3t_{1/2}$	125,000
$4t_{1/2}$	62,500
$5t_{1/2}$	31,250
$6t_{1/2}$	15,625
$7t_{1/2}$	7,813
$8t_{1/2}$	3,906
$9t_{1/2}$	1,953
$10t_{1/2}$	977

Radioactive decay reduces the number of radioactive nuclei over time. In one half-life $t_{1/2}$, the number decreases to half of its original value. Half of what remains decay in the next half-life, and half of those in the next, and so on. This is an exponential decay, as seen in the graph of the number of nuclei present as a function of time.

There is a tremendous range in the half-lives of various nuclides, from as short as 10^{-23} s for the most unstable, to more than 10^{16} y for the least unstable, or about 46 orders of magnitude. Nuclides with the shortest half-lives are those for which the nuclear forces are least attractive, an indication of the extent to which the nuclear force can depend on the particular combination of neutrons and protons. The concept of half-life is applicable to other subatomic particles, as will be discussed in Particle Physics. It is also applicable to the decay of excited states in atoms and nuclei. The following equation gives the quantitative relationship between the original number of nuclei present at time zero (N_0) and the number (N) at a later time t:

Equation:

$$N = N_0 e^{-\lambda t},$$

where $e = 2.71828...$ is the base of the natural logarithm, and λ is the **decay constant** for the nuclide. The shorter the half-life, the larger is the value of λ, and the faster the exponential $e^{-\lambda t}$ decreases with time. The relationship between the decay constant λ and the half-life $t_{1/2}$ is

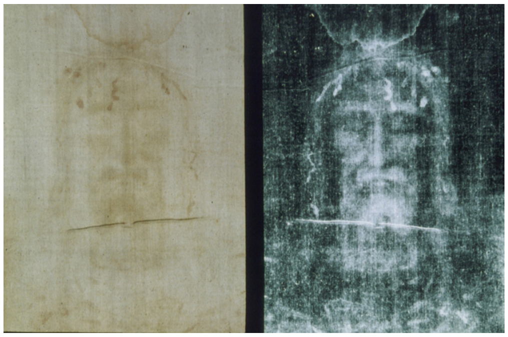
Equation:

$$\lambda = \frac{\ln(2)}{t_{1/2}} \approx \frac{0.693}{t_{1/2}}.$$

To see how the number of nuclei declines to half its original value in one half-life, let $t = t_{1/2}$ in the exponential in the equation $N = N_0 e^{-\lambda t}$. This gives $N = N_0 e^{-\lambda t} = N_0 e^{-0.693} = 0.500 N_0$. For integral numbers of half-lives, you can just divide the original number by 2 over and over again, rather than using the exponential relationship. For example, if ten half-lives have passed, we divide N by 2 ten times. This reduces it to $N/1024$. For an arbitrary time, not just a multiple of the half-life, the exponential relationship must be used.

Radioactive dating is a clever use of naturally occurring radioactivity. Its most famous application is **carbon-14 dating**. Carbon-14 has a half-life of 5730 years and is produced in a nuclear reaction induced when solar neutrinos strike ^{14}N in the atmosphere. Radioactive carbon has the same chemistry as stable carbon, and so it mixes into the ecosphere, where it is consumed and becomes part of every living organism. Carbon-14 has an abundance of 1.3 parts per trillion of normal carbon. Thus, if you know the number of carbon nuclei in an object (perhaps determined by mass and Avogadro’s number), you multiply that number by 1.3×10^{-12} to find the number of ^{14}C nuclei in the object. When an organism dies, carbon exchange with the environment ceases, and ^{14}C is not replenished as it decays. By comparing the abundance of ^{14}C in an artifact, such as mummy wrappings, with the normal abundance in living tissue, it is possible to determine the artifact's age (or time since death). Carbon-14 dating can be used for biological tissues as old as 50 or 60 thousand years, but is most accurate for younger samples, since the abundance of ^{14}C nuclei in them is greater. Very old biological materials contain no ^{14}C at all. There are instances in which the date of an artifact can be determined by other means, such as historical knowledge or tree-ring counting. These cross-references have confirmed the validity of carbon-14 dating and permitted us to calibrate the technique as well. Carbon-14 dating revolutionized parts of archaeology and is of such importance that it earned the 1960 Nobel Prize in chemistry for its developer, the American chemist Willard Libby (1908–1980).

One of the most famous cases of carbon-14 dating involves the Shroud of Turin, a long piece of fabric purported to be the burial shroud of Jesus (see [link]). This relic was first displayed in Turin in 1354 and was denounced as a fraud at that time by a French bishop. Its remarkable negative imprint of an apparently crucified body resembles the then-accepted image of Jesus, and so the shroud was never disregarded completely and remained controversial over the centuries. Carbon-14 dating was not performed on the shroud until 1988, when the process had been refined to the point where only a small amount of material needed to be destroyed. Samples were tested at three independent laboratories, each being given four pieces of cloth, with only one unidentified piece from the shroud, to avoid prejudice. All three laboratories found samples of the shroud contain 92% of the ^{14}C found in living tissues, allowing the shroud to be dated (see [link]).



Part of the Shroud of Turin, which shows a remarkable negative imprint likeness of Jesus complete with evidence of crucifixion wounds. The shroud first surfaced in the 14th century and was only recently carbon-14 dated. It has not been determined how the image was placed on the material. (credit: Butko, Wikimedia Commons)

Example:

How Old Is the Shroud of Turin?

Calculate the age of the Shroud of Turin given that the amount of ^{14}C found in it is 92% of that in living tissue.

Strategy

Knowing that 92% of the ^{14}C remains means that $N/N_0 = 0.92$. Therefore, the equation $N = N_0 e^{-\lambda t}$ can be used to find λt. We also know that the half-life of ^{14}C is 5730 y, and so once λt is known, we can use the equation $\lambda = \frac{0.693}{t_{1/2}}$ to find λ and then find t as requested. Here, we postulate that the decrease in ^{14}C is solely due to nuclear decay.

Solution

Solving the equation $N = N_0 e^{-\lambda t}$ for N/N_0 gives

Equation:

$$\frac{N}{N_0} = e^{-\lambda t}.$$

Thus,
Equation:

$$0.92 = e^{-\lambda t}.$$

Taking the natural logarithm of both sides of the equation yields
Equation:

$$\ln 0.92 = -\lambda t$$

so that
Equation:

$$-0.0834 = -\lambda t.$$

Rearranging to isolate t gives
Equation:

$$t = \frac{0.0834}{\lambda}.$$

Now, the equation $\lambda = \frac{0.693}{t_{1/2}}$ can be used to find λ for ^{14}C. Solving for λ and substituting the known half-life gives
Equation:

$$\lambda = \frac{0.693}{t_{1/2}} = \frac{0.693}{5730\ \text{y}}.$$

We enter this value into the previous equation to find t:
Equation:

$$t = \frac{0.0834}{\frac{0.693}{5730\ \text{y}}} = 690\ \text{y}.$$

Discussion
This dates the material in the shroud to 1988–690 = a.d. 1300. Our calculation is only accurate to two digits, so that the year is rounded to 1300. The values obtained at the three independent laboratories gave a

weighted average date of a.d. 1320 ± 60. The uncertainty is typical of carbon-14 dating and is due to the small amount of $^{14}\mathrm{C}$ in living tissues, the amount of material available, and experimental uncertainties (reduced by having three independent measurements). It is meaningful that the date of the shroud is consistent with the first record of its existence and inconsistent with the period in which Jesus lived.

There are other forms of radioactive dating. Rocks, for example, can sometimes be dated based on the decay of $^{238}\mathrm{U}$. The decay series for $^{238}\mathrm{U}$ ends with $^{206}\mathrm{Pb}$, so that the ratio of these nuclides in a rock is an indication of how long it has been since the rock solidified. The original composition of the rock, such as the absence of lead, must be known with some confidence. However, as with carbon-14 dating, the technique can be verified by a consistent body of knowledge. Since $^{238}\mathrm{U}$ has a half-life of 4.5×10^9 y, it is useful for dating only very old materials, showing, for example, that the oldest rocks on Earth solidified about 3.5×10^9 years ago.

Activity, the Rate of Decay

What do we mean when we say a source is highly radioactive? Generally, this means the number of decays per unit time is very high. We define **activity** R to be the **rate of decay** expressed in decays per unit time. In equation form, this is

Equation:

$$R = \frac{\Delta N}{\Delta t}$$

where ΔN is the number of decays that occur in time Δt. The SI unit for activity is one decay per second and is given the name **becquerel** (Bq) in honor of the discoverer of radioactivity. That is,

Equation:

$$1\text{ Bq} = 1\text{ decay/s}.$$

Activity R is often expressed in other units, such as decays per minute or decays per year. One of the most common units for activity is the **curie** (Ci), defined to be the activity of 1 g of ^{226}Ra, in honor of Marie Curie's work with radium. The definition of curie is

Equation:

$$1\text{ Ci} = 3.70 \times 10^{10}\text{ Bq},$$

or 3.70×10^{10} decays per second. A curie is a large unit of activity, while a becquerel is a relatively small unit. $1\text{ MBq} = 100\text{ microcuries }(\mu\text{Ci})$. In countries like Australia and New Zealand that adhere more to SI units, most radioactive sources, such as those used in medical diagnostics or in physics laboratories, are labeled in Bq or megabecquerel (MBq).

Intuitively, you would expect the activity of a source to depend on two things: the amount of the radioactive substance present, and its half-life. The greater the number of radioactive nuclei present in the sample, the more will decay per unit of time. The shorter the half-life, the more decays per unit time, for a given number of nuclei. So activity R should be proportional to the number of radioactive nuclei, N, and inversely proportional to their half-life, $t_{1/2}$. In fact, your intuition is correct. It can be shown that the activity of a source is

Equation:

$$R = \frac{0.693N}{t_{1/2}}$$

where N is the number of radioactive nuclei present, having half-life $t_{1/2}$. This relationship is useful in a variety of calculations, as the next two examples illustrate.

Example:

How Great Is the ^{14}C Activity in Living Tissue?

Calculate the activity due to ^{14}C in 1.00 kg of carbon found in a living organism. Express the activity in units of Bq and Ci.

Strategy

To find the activity R using the equation $R = \frac{0.693N}{t_{1/2}}$, we must know N and $t_{1/2}$. The half-life of ^{14}C can be found in Appendix B, and was stated above as 5730 y. To find N, we first find the number of ^{12}C nuclei in 1.00 kg of carbon using the concept of a mole. As indicated, we then multiply by 1.3×10^{-12} (the abundance of ^{14}C in a carbon sample from a living organism) to get the number of ^{14}C nuclei in a living organism.

Solution

One mole of carbon has a mass of 12.0 g, since it is nearly pure ^{12}C. (A mole has a mass in grams equal in magnitude to A found in the periodic table.) Thus the number of carbon nuclei in a kilogram is

Equation:

$$N(^{12}\text{C}) = \frac{6.02 \times 10^{23}\ \text{mol}^{-1}}{12.0\ \text{g/mol}} \times (1000\ \text{g}) = 5.02 \times 10^{25}.$$

So the number of ^{14}C nuclei in 1 kg of carbon is

Equation:

$$N(^{14}\text{C}) = (5.02 \times 10^{25})(1.3 \times 10^{-12}) = 6.52 \times 10^{13}.$$

Now the activity R is found using the equation $R = \frac{0.693N}{t_{1/2}}$.

Entering known values gives

Equation:

$$R = \frac{0.693(6.52 \times 10^{13})}{5730\ \text{y}} = 7.89 \times 10^{9}\ \text{y}^{-1},$$

or 7.89×10^{9} decays per year. To convert this to the unit Bq, we simply convert years to seconds. Thus,

Equation:

$$R = (7.89 \times 10^{9}\text{ y}^{-1})\frac{1.00\text{ y}}{3.16 \times 10^{7}\text{ s}} = 250\text{ Bq},$$

or 250 decays per second. To express R in curies, we use the definition of a curie,

Equation:

$$R = \frac{250\text{ Bq}}{3.7 \times 10^{10}\text{ Bq/Ci}} = 6.76 \times 10^{-9}\text{ Ci.}$$

Thus,

Equation:

$$R = 6.76\text{ nCi.}$$

Discussion

Our own bodies contain kilograms of carbon, and it is intriguing to think there are hundreds of ^{14}C decays per second taking place in us. Carbon-14 and other naturally occurring radioactive substances in our bodies contribute to the background radiation we receive. The small number of decays per second found for a kilogram of carbon in this example gives you some idea of how difficult it is to detect ^{14}C in a small sample of material. If there are 250 decays per second in a kilogram, then there are 0.25 decays per second in a gram of carbon in living tissue. To observe this, you must be able to distinguish decays from other forms of radiation, in order to reduce background noise. This becomes more difficult with an old tissue sample, since it contains less ^{14}C, and for samples more than 50 thousand years old, it is impossible.

Human-made (or artificial) radioactivity has been produced for decades and has many uses. Some of these include medical therapy for cancer, medical imaging and diagnostics, and food preservation by irradiation. Many applications as well as the biological effects of radiation are explored in Medical Applications of Nuclear Physics, but it is clear that radiation is hazardous. A number of tragic examples of this exist, one of the most disastrous being the meltdown and fire at the Chernobyl reactor complex in the Ukraine (see [link]). Several radioactive isotopes were released in huge quantities, contaminating many thousands of square kilometers and directly affecting hundreds of thousands of people. The most significant releases were of ^{131}I, ^{90}Sr, ^{137}Cs, ^{239}Pu, ^{238}U, and ^{235}U. Estimates are that the total amount of radiation released was about 100 million curies.

Human and Medical Applications

The Chernobyl reactor. More than 100 people died soon after its meltdown, and there will be thousands of deaths from radiation-induced cancer in the future. While the accident was due to a series of human errors, the cleanup efforts were heroic. Most of the immediate fatalities were firefighters and reactor personnel. (credit: Elena Filatova)

Example:

What Mass of ^{137}Cs Escaped Chernobyl?

It is estimated that the Chernobyl disaster released 6.0 MCi of ^{137}Cs into the environment. Calculate the mass of ^{137}Cs released.

Strategy

We can calculate the mass released using Avogadro's number and the concept of a mole if we can first find the number of nuclei N released. Since the activity R is given, and the half-life of ^{137}Cs is found in Appendix B to be 30.2 y, we can use the equation $R=\frac{0.693N}{t_{1/2}}$ to find N.

Solution

Solving the equation $R=\frac{0.693N}{t_{1/2}}$ for N gives

Equation:

$$N=\frac{Rt_{1/2}}{0.693}.$$

Entering the given values yields

Equation:

$$N=\frac{(6.0\ \text{MCi})(30.2\ \text{y})}{0.693}.$$

Converting curies to becquerels and years to seconds, we get

Equation:

$$\begin{aligned} N &= \frac{(6.0\times10^{6}\ \text{Ci})(3.7\times10^{10}\ \text{Bq/Ci})(30.2\ \text{y})(3.16\times10^{7}\ \text{s/y})}{0.693} \\ &= 3.1\times10^{26}. \end{aligned}$$

One mole of a nuclide ^{A}X has a mass of A grams, so that one mole of ^{137}Cs has a mass of 137 g. A mole has 6.02×10^{23} nuclei. Thus the mass of ^{137}Cs released was

Equation:

$$\begin{aligned} m &= \left(\frac{137\ \text{g}}{6.02\times10^{23}}\right)(3.1\times10^{26})=70\times10^{3}\ \text{g} \\ &= 70\ \text{kg}. \end{aligned}$$

Discussion

While 70 kg of material may not be a very large mass compared to the amount of fuel in a power plant, it is extremely radioactive, since it only has a 30-year half-life. Six megacuries (6.0 MCi) is an extraordinary amount of activity but is only a fraction of what is produced in nuclear reactors. Similar amounts of the other isotopes were also released at Chernobyl. Although the chances of such a disaster may have seemed small, the consequences were extremely severe, requiring greater caution than was used. More will be said about safe reactor design in the next chapter, but it should be noted that Western reactors have a fundamentally safer design.

Activity R decreases in time, going to half its original value in one half-life, then to one-fourth its original value in the next half-life, and so on. Since $R = \frac{0.693N}{t_{1/2}}$, the activity decreases as the number of radioactive nuclei decreases. The equation for R as a function of time is found by combining the equations $N = N_0 e^{-\lambda t}$ and $R = \frac{0.693N}{t_{1/2}}$, yielding

Equation:

$$R = R_0 e^{-\lambda t},$$

where R_0 is the activity at $t = 0$. This equation shows exponential decay of radioactive nuclei. For example, if a source originally has a 1.00-mCi activity, it declines to 0.500 mCi in one half-life, to 0.250 mCi in two half-lives, to 0.125 mCi in three half-lives, and so on. For times other than whole half-lives, the equation $R = R_0 e^{-\lambda t}$ must be used to find R.

Note:

PhET Explorations: Alpha Decay

Watch alpha particles escape from a polonium nucleus, causing radioactive alpha decay. See how random decay times relate to the half life.

Alpha Decay

Section Summary

- Half-life $t_{1/2}$ is the time in which there is a 50% chance that a nucleus will decay. The number of nuclei N as a function of time is
  Equation:
  $$N = N_0 e^{-\lambda t},$$
  where N_0 is the number present at $t = 0$, and λ is the decay constant, related to the half-life by
  Equation:
  $$\lambda = \frac{0.693}{t_{1/2}}.$$
- One of the applications of radioactive decay is radioactive dating, in which the age of a material is determined by the amount of radioactive decay that occurs. The rate of decay is called the activity R:
  Equation:
  $$R = \frac{\Delta N}{\Delta t}.$$
- The SI unit for R is the becquerel (Bq), defined by
  Equation:
  $$1\text{ Bq} = 1\text{ decay/s}.$$
- R is also expressed in terms of curies (Ci), where

Equation:

$$1\text{ Ci} = 3.70 \times 10^{10}\text{ Bq}.$$

- The activity R of a source is related to N and $t_{1/2}$ by
  Equation:

  $$R = \frac{0.693N}{t_{1/2}}.$$

- Since N has an exponential behavior as in the equation $N = N_0 e^{-\lambda t}$, the activity also has an exponential behavior, given by
  Equation:

  $$R = R_0 e^{-\lambda t},$$

  where R_0 is the activity at $t = 0$.

Conceptual Questions

Exercise:

Problem:

In a 3×10^9-year-old rock that originally contained some ^{238}U, which has a half-life of 4.5×10^9 years, we expect to find some ^{238}U remaining in it. Why are ^{226}Ra, ^{222}Rn, and ^{210}Po also found in such a rock, even though they have much shorter half-lives (1600 years, 3.8 days, and 138 days, respectively)?

Exercise:

Problem:

Does the number of radioactive nuclei in a sample decrease to *exactly* half its original value in one half-life? Explain in terms of the statistical nature of radioactive decay.

Exercise:

Problem:

Radioactivity depends on the nucleus and not the atom or its chemical state. Why, then, is one kilogram of uranium more radioactive than one kilogram of uranium hexafluoride?

Exercise:

Problem:

Explain how a bound system can have less mass than its components. Why is this not observed classically, say for a building made of bricks?

Exercise:

Problem:

Spontaneous radioactive decay occurs only when the decay products have less mass than the parent, and it tends to produce a daughter that is more stable than the parent. Explain how this is related to the fact that more tightly bound nuclei are more stable. (Consider the binding energy per nucleon.)

Exercise:

Problem:

To obtain the most precise value of BE from the equation $\text{BE}=\left[ZM\left({}^{1}\text{H}\right)+Nm_{n}\right]c^{2}-m\left({}^{A}\text{X}\right)c^{2}$, we should take into account the binding energy of the electrons in the neutral atoms. Will doing this produce a larger or smaller value for BE? Why is this effect usually negligible?

Exercise:

Problem:

How does the finite range of the nuclear force relate to the fact that BE/A is greatest for A near 60?

Problems & Exercises

Data from the appendices and the periodic table may be needed for these problems.

Exercise:

Problem:

An old campfire is uncovered during an archaeological dig. Its charcoal is found to contain less than 1/1000 the normal amount of ^{14}C. Estimate the minimum age of the charcoal, noting that $2^{10} = 1024$.

Solution:

57,300 y

Exercise:

Problem:

A ^{60}Co source is labeled 4.00 mCi, but its present activity is found to be 1.85×10^7 Bq. (a) What is the present activity in mCi? (b) How long ago did it actually have a 4.00-mCi activity?

Exercise:

Problem:

(a) Calculate the activity R in curies of 1.00 g of ^{226}Ra. (b) Discuss why your answer is not exactly 1.00 Ci, given that the curie was originally supposed to be exactly the activity of a gram of radium.

Solution:

(a) 0.988 Ci

(b) The half-life of ^{226}Ra is now better known.

Exercise:

Problem:

Show that the activity of the ^{14}C in 1.00 g of ^{12}C found in living tissue is 0.250 Bq.

Exercise:

Problem:

Mantles for gas lanterns contain thorium, because it forms an oxide that can survive being heated to incandescence for long periods of time. Natural thorium is almost 100% ^{232}Th, with a half-life of 1.405×10^{10} y. If an average lantern mantle contains 300 mg of thorium, what is its activity?

Solution:

1.22×10^{3} Bq

Exercise:

Problem:

Cow’s milk produced near nuclear reactors can be tested for as little as 1.00 pCi of ^{131}I per liter, to check for possible reactor leakage. What mass of ^{131}I has this activity?

Exercise:

Problem:

(a) Natural potassium contains ^{40}K, which has a half-life of 1.277×10^{9} y. What mass of ^{40}K in a person would have a decay rate of 4140 Bq? (b) What is the fraction of ^{40}K in natural potassium, given that the person has 140 g in his body? (These numbers are typical for a 70-kg adult.)

Solution:

(a) 16.0 mg

(b) 0.0114%

Exercise:

Problem:

There is more than one isotope of natural uranium. If a researcher isolates 1.00 mg of the relatively scarce ^{235}U and finds this mass to have an activity of 80.0 Bq, what is its half-life in years?

Exercise:

Problem:

^{50}V has one of the longest known radioactive half-lives. In a difficult experiment, a researcher found that the activity of 1.00 kg of ^{50}V is 1.75 Bq. What is the half-life in years?

Solution:

1.48×10^{17} y

Exercise:

Problem:

You can sometimes find deep red crystal vases in antique stores, called uranium glass because their color was produced by doping the glass with uranium. Look up the natural isotopes of uranium and their half-lives, and calculate the activity of such a vase assuming it has 2.00 g of uranium in it. Neglect the activity of any daughter nuclides.

Exercise:

Problem:

A tree falls in a forest. How many years must pass before the ^{14}C activity in 1.00 g of the tree’s carbon drops to 1.00 decay per hour?

Solution:

5.6×10^{4} y

Exercise:

Problem:

What fraction of the ^{40}K that was on Earth when it formed 4.5×10^9 years ago is left today?

Exercise:

Problem:

A 5000-Ci ^{60}Co source used for cancer therapy is considered too weak to be useful when its activity falls to 3500 Ci. How long after its manufacture does this happen?

Solution:

2.71 y

Exercise:

Problem:

Natural uranium is 0.7200% ^{235}U and 99.27% ^{238}U. What were the percentages of ^{235}U and ^{238}U in natural uranium when Earth formed 4.5×10^9 years ago?

Exercise:

Problem:

The β^- particles emitted in the decay of ^{3}H (tritium) interact with matter to create light in a glow-in-the-dark exit sign. At the time of manufacture, such a sign contains 15.0 Ci of ^{3}H. (a) What is the mass of the tritium? (b) What is its activity 5.00 y after manufacture?

Solution:

(a) 1.56 mg

(b) 11.3 Ci

Exercise:

Problem:

World War II aircraft had instruments with glowing radium-painted dials (see [link]). The activity of one such instrument was 1.0×10^5 Bq when new. (a) What mass of ^{226}Ra was present? (b) After some years, the phosphors on the dials deteriorated chemically, but the radium did not escape. What is the activity of this instrument 57.0 years after it was made?

Exercise:

Problem:

(a) The ^{210}Po source used in a physics laboratory is labeled as having an activity of $1.0\ \mu\text{Ci}$ on the date it was prepared. A student measures the radioactivity of this source with a Geiger counter and observes 1500 counts per minute. She notices that the source was prepared 120 days before her lab. What fraction of the decays is she observing with her apparatus? (b) Identify some of the reasons that only a fraction of the α s emitted are observed by the detector.

Solution:

(a) 1.23×10^{-3}

(b) Only part of the emitted radiation goes in the direction of the detector. Only a fraction of that causes a response in the detector. Some of the emitted radiation (mostly α particles) is observed within the source. Some is absorbed within the source, some is absorbed by the detector, and some does not penetrate the detector.

Exercise:

Problem:

Armor-piercing shells with depleted uranium cores are fired by aircraft at tanks. (The high density of the uranium makes them effective.) The uranium is called depleted because it has had its $^{235}\mathrm{U}$ removed for reactor use and is nearly pure $^{238}\mathrm{U}$. Depleted uranium has been erroneously called non-radioactive. To demonstrate that this is wrong: (a) Calculate the activity of 60.0 g of pure $^{238}\mathrm{U}$. (b) Calculate the activity of 60.0 g of natural uranium, neglecting the $^{234}\mathrm{U}$ and all daughter nuclides.

Exercise:

Problem:

The ceramic glaze on a red-orange Fiestaware plate is $\mathrm{U_2O_3}$ and contains 50.0 grams of $^{238}\mathrm{U}$, but very little $^{235}\mathrm{U}$. (a) What is the activity of the plate? (b) Calculate the total energy that will be released by the $^{238}\mathrm{U}$ decay. (c) If energy is worth 12.0 cents per $\mathrm{kW \cdot h}$, what is the monetary value of the energy emitted? (These plates went out of production some 30 years ago, but are still available as collectibles.)

Solution:

(a) 1.68×10^{-5} Ci

(b) 8.65×10^{10} J

(c) $\$2.9 \times 10^{3}$

Exercise:

Problem:

Large amounts of depleted uranium (^{238}U) are available as a by-product of uranium processing for reactor fuel and weapons. Uranium is very dense and makes good counter weights for aircraft. Suppose you have a 4000-kg block of ^{238}U. (a) Find its activity. (b) How many calories per day are generated by thermalization of the decay energy? (c) Do you think you could detect this as heat? Explain.

Exercise:

Problem:

The *Galileo* space probe was launched on its long journey past several planets in 1989, with an ultimate goal of Jupiter. Its power source is 11.0 kg of ^{238}Pu, a by-product of nuclear weapons plutonium production. Electrical energy is generated thermoelectrically from the heat produced when the 5.59-MeV α particles emitted in each decay crash to a halt inside the plutonium and its shielding. The half-life of ^{238}Pu is 87.7 years. (a) What was the original activity of the ^{238}Pu in becquerel? (b) What power was emitted in kilowatts? (c) What power was emitted 12.0 y after launch? You may neglect any extra energy from daughter nuclides and any losses from escaping γ rays.

Solution:

(a) $6.97 \times 10^{15}\ \text{Bq}$

(b) 6.24 kW

(c) 5.67 kW

Exercise:

Problem: Construct Your Own Problem

Consider the generation of electricity by a radioactive isotope in a space probe, such as described in [link]. Construct a problem in which you calculate the mass of a radioactive isotope you need in order to supply power for a long space flight. Among the things to consider are the isotope chosen, its half-life and decay energy, the power needs of the probe and the length of the flight.

Exercise:

Problem: Unreasonable Results

A nuclear physicist finds $1.0\ \mu\text{g}$ of ^{236}U in a piece of uranium ore and assumes it is primordial since its half-life is $2.3 \times 10^{7}\ \text{y}$. (a) Calculate the amount of ^{236}Uthat would had to have been on Earth when it formed $4.5 \times 10^{9}\ \text{y}$ ago for $1.0\ \mu\text{g}$ to be left today. (b) What is unreasonable about this result? (c) What assumption is responsible?

Exercise:

Problem: Unreasonable Results

(a) Repeat [link] but include the 0.0055% natural abundance of ^{234}U with its $2.45 \times 10^{5}\ \text{y}$ half-life. (b) What is unreasonable about this result? (c) What assumption is responsible? (d) Where does the ^{234}U come from if it is not primordial?

Exercise:

Problem: Unreasonable Results

The manufacturer of a smoke alarm decides that the smallest current of α radiation he can detect is $1.00\ \mu\text{A}$. (a) Find the activity in curies of an α emitter that produces a $1.00\ \mu\text{A}$ current of α particles. (b) What is unreasonable about this result? (c) What assumption is responsible?

Solution:

(a) 84.5 Ci

(b) An extremely large activity, many orders of magnitude greater than permitted for home use.

(c) The assumption of $1.00\ \mu\text{A}$ is unreasonably large. Other methods can detect much smaller decay rates.

Glossary

becquerel
SI unit for rate of decay of a radioactive material

half-life
the time in which there is a 50% chance that a nucleus will decay

radioactive dating
an application of radioactive decay in which the age of a material is determined by the amount of radioactivity of a particular type that occurs

decay constant
quantity that is inversely proportional to the half-life and that is used in equation for number of nuclei as a function of time

carbon-14 dating
a radioactive dating technique based on the radioactivity of carbon-14

activity
the rate of decay for radioactive nuclides

rate of decay
the number of radioactive events per unit time

curie
the activity of 1g of ^{226}Ra, equal to $3.70 \times 10^{10}\ \text{Bq}$

Binding Energy

- Define and discuss binding energy.
- Calculate the binding energy per nucleon of a particle.

The more tightly bound a system is, the stronger the forces that hold it together and the greater the energy required to pull it apart. We can therefore learn about nuclear forces by examining how tightly bound the nuclei are. We define the **binding energy** (BE) of a nucleus to be *the energy required to completely disassemble it into separate protons and neutrons*. We can determine the BE of a nucleus from its rest mass. The two are connected through Einstein's famous relationship $E=(\Delta m)c^2$. A bound system has a *smaller* mass than its separate constituents; the more tightly the nucleons are bound together, the smaller the mass of the nucleus.

Imagine pulling a nuclide apart as illustrated in [link]. Work done to overcome the nuclear forces holding the nucleus together puts energy into the system. By definition, the energy input equals the binding energy BE. The pieces are at rest when separated, and so the energy put into them increases their total rest mass compared with what it was when they were glued together as a nucleus. That mass increase is thus $\Delta m=\text{BE}/c^2$. This difference in mass is known as *mass defect*. It implies that the mass of the nucleus is less than the sum of the masses of its constituent protons and neutrons. A nuclide ${}^{A}\text{X}$ has Z protons and N neutrons, so that the difference in mass is

Equation:

$$\Delta m=(Zm_p+\text{Nm}_n)-m_{\text{tot}}.$$

Thus,

Equation:

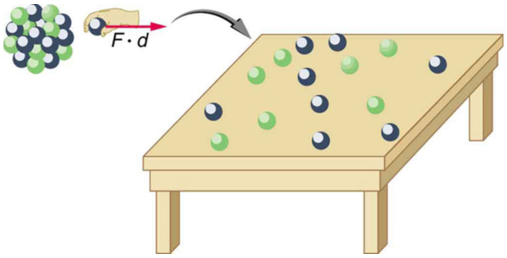
$$\text{BE}=(\Delta m)c^2=[(\text{Zm}_p+\text{Nm}_n)-m_{\text{tot}}]c^2,$$

where m_{tot} is the mass of the nuclide ${}^{A}\text{X}$, m_p is the mass of a proton, and m_n is the mass of a neutron. Traditionally, we deal with the masses of neutral atoms. To get atomic masses into the last equation, we first add Z electrons to m_{tot}, which gives $m\left({}^{A}\text{X}\right)$, the atomic mass of the nuclide. We then add Z electrons to the Z protons, which gives $\text{Zm}\left({}^{1}\text{H}\right)$, or Z times the mass of a hydrogen atom. Thus the binding energy of a nuclide ${}^{A}\text{X}$ is

Equation:

$$\text{BE} = \left\{[Zm({}^{1}\text{H}) + Nm_{n}] - m({}^{A}X)\right\}c^{2}.$$

The atomic masses can be found in Appendix A, most conveniently expressed in unified atomic mass units u ($1\ \text{u} = 931.5\ \text{MeV}/c^{2}$). BE is thus calculated from known atomic masses.



Work done to pull a nucleus apart into its constituent protons and neutrons increases the mass of the system. The work to disassemble the nucleus equals its binding energy BE. A bound system has less mass than the sum of its parts, especially noticeable in the nuclei, where forces and energies are very large.

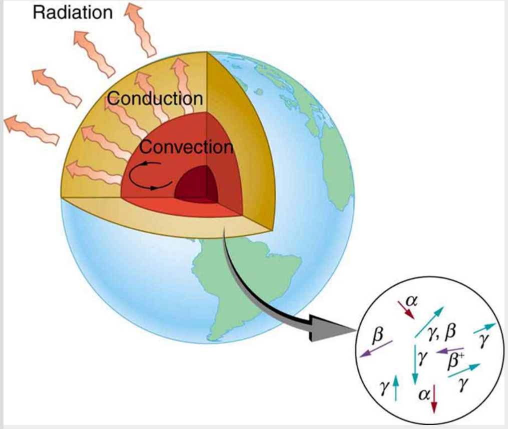
Note:

Things Great and Small

Nuclear Decay Helps Explain Earth’s Hot Interior

A puzzle created by radioactive dating of rocks is resolved by radioactive heating of Earth’s interior. This intriguing story is another example of how small-scale physics can explain large-scale phenomena.

Radioactive dating plays a role in determining the approximate age of the Earth. The oldest rocks on Earth solidified about 3.5×10^9 years ago—a number determined by uranium-238 dating. These rocks could only have solidified once the surface of the Earth had cooled sufficiently. The temperature of the Earth at formation can be estimated based on gravitational potential energy of the assemblage of pieces being converted to thermal energy. Using heat transfer concepts discussed in Thermodynamics it is then possible to calculate how long it would take for the surface to cool to rock-formation temperatures. The result is about 10^9 years. The first rocks formed have been solid for 3.5×10^9 years, so that the age of the Earth is approximately 4.5×10^9 years. There is a large body of other types of evidence (both Earth-bound and solar system characteristics are used) that supports this age. The puzzle is that, given its age and initial temperature, the center of the Earth should be much cooler than it is today (see [link]).



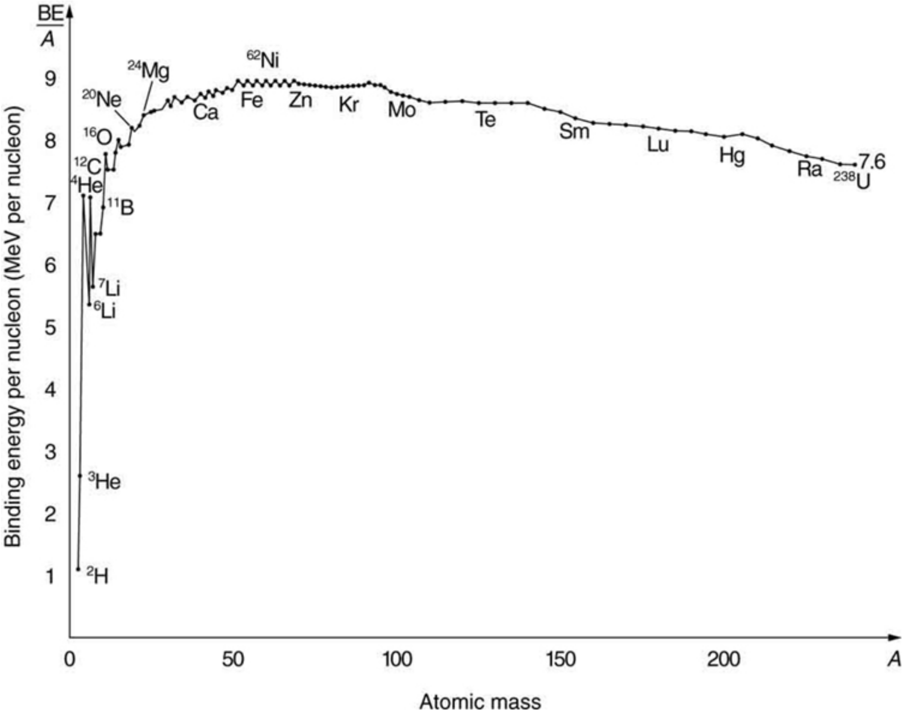
The center of the Earth cools by well-known heat transfer methods. Convection in the liquid regions and conduction move thermal energy to the surface, where it radiates into cold, dark space. Given the age of the Earth and its initial temperature, it should have cooled to a lower temperature by now. The blowup shows that nuclear decay releases energy in the Earth’s interior. This energy has slowed the cooling process and is responsible for the interior still being molten.

We know from seismic waves produced by earthquakes that parts of the interior of the Earth are liquid. Shear or transverse waves cannot travel through a liquid and are not transmitted through the Earth’s core. Yet compression or longitudinal waves can pass through a liquid and do go through the core. From this information, the temperature of the interior can be estimated. As noticed, the interior should have cooled more from its initial temperature in the 4.5×10^9 years since its formation. In fact, it should have taken no more than about 10^9 years to cool to its present temperature. What is keeping it hot? The answer seems to be radioactive decay of primordial elements that were part of the material that formed the Earth (see the blowup in [link]).

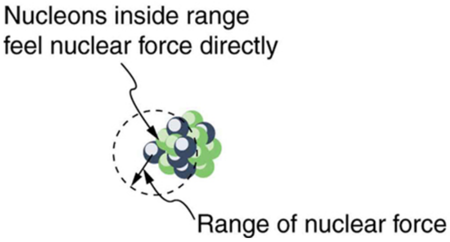
Nuclides such as ^{238}U and ^{40}K have half-lives similar to or longer than the age of the Earth, and their decay still contributes energy to the interior. Some of the primordial radioactive nuclides have unstable decay products that also release energy— ^{238}U has a long decay chain of these. Further, there were more of these primordial radioactive nuclides early in the life of the Earth, and thus the activity and energy contributed were greater then (perhaps by an order of magnitude). The amount of power created by these decays per cubic meter is very small. However, since a huge volume of material lies deep below the surface, this relatively small amount of energy cannot escape quickly. The power produced near the surface has much less distance to go to escape and has a negligible effect on surface temperatures.

A final effect of this trapped radiation merits mention. Alpha decay produces helium nuclei, which form helium atoms when they are stopped and capture electrons. Most of the helium on Earth is obtained from wells and is produced in this manner. Any helium in the atmosphere will escape in geologically short times because of its high thermal velocity.

What patterns and insights are gained from an examination of the binding energy of various nuclides? First, we find that BE is approximately proportional to the number of nucleons A in any nucleus. About twice as much energy is needed to pull apart a nucleus like ^{24}Mg compared with pulling apart ^{12}C, for example. To help us look at other effects, we divide BE by A and consider the **binding energy per nucleon**, BE/A. The graph of BE/A in [link] reveals some very interesting aspects of nuclei. We see that the binding energy per nucleon averages about 8 MeV, but is lower for both the lightest and heaviest nuclei. This overall trend, in which nuclei with A equal to about 60 have the greatest BE/A and are thus the most tightly bound, is due to the combined characteristics of the attractive nuclear forces and the repulsive Coulomb force. It is especially important to note two things—the strong nuclear force is about 100 times stronger than the Coulomb force, *and* the nuclear forces are shorter in range compared to the Coulomb force. So, for low-mass nuclei, the nuclear attraction dominates and each added nucleon forms bonds with all others, causing progressively heavier nuclei to have progressively greater values of BE/A. This continues up to $A \approx 60$, roughly corresponding to the mass number of iron. Beyond that, new nucleons added to a nucleus will be too far from some others to feel their nuclear attraction. Added protons, however, feel the repulsion of all other protons, since the Coulomb force is longer in range. Coulomb repulsion grows for progressively heavier nuclei, but nuclear attraction remains about the same, and so BE/A becomes smaller. This is why stable nuclei heavier than $A \approx 40$ have more neutrons than protons. Coulomb repulsion is reduced by having more neutrons to keep the protons farther apart (see [link]).



A graph of average binding energy per nucleon, BE/A, for stable nuclei. The most tightly bound nuclei are those with A near 60, where the attractive nuclear force has its greatest effect. At higher A s, the Coulomb repulsion progressively reduces the binding energy per nucleon, because the nuclear force is short ranged. The spikes on the curve are very tightly bound nuclides and indicate shell closures.



The nuclear force is

attractive and stronger than the Coulomb force, but it is short ranged. In low-mass nuclei, each nucleon feels the nuclear attraction of all others. In larger nuclei, the range of the nuclear force, shown for a single nucleon, is smaller than the size of the nucleus, but the Coulomb repulsion from all protons reaches all others. If the nucleus is large enough, the Coulomb repulsion can add to overcome the nuclear attraction.

There are some noticeable spikes on the BE/A graph, which represent particularly tightly bound nuclei. These spikes reveal further details of nuclear forces, such as confirming that closed-shell nuclei (those with magic numbers of protons or neutrons or both) are more tightly bound. The spikes also indicate that some nuclei with even numbers for Z and N, and with $Z = N$, are exceptionally tightly bound. This finding can be correlated with some of the cosmic abundances of the elements. The most common elements in the universe, as determined by observations of atomic spectra from outer space, are hydrogen, followed by $^{4}\mathrm{He}$, with much smaller amounts of $^{12}\mathrm{C}$ and other elements. It should be noted that the heavier elements are created in supernova explosions, while the lighter ones are produced by nuclear fusion during the normal life cycles of stars, as will be discussed in subsequent chapters. The most common elements have the most tightly bound nuclei. It is also no accident that one of the most tightly bound light nuclei is ${}^{4}\text{He}$, emitted in α decay.

Example:

What Is BE/A for an Alpha Particle?

Calculate the binding energy per nucleon of ${}^{4}\text{He}$, the α particle.

Strategy

To find BE/A, we first find BE using the Equation $\text{BE} = \{[\text{Zm}({}^{1}\text{H}) + \text{Nm}_{n}] - m({}^{A}\text{X})\}c^{2}$ and then divide by A. This is straightforward once we have looked up the appropriate atomic masses in Appendix A.

Solution

The binding energy for a nucleus is given by the equation

Equation:

$$\text{BE} = \{[\text{Zm}({}^{1}\text{H}) + \text{Nm}_{n}] - m({}^{A}\text{X})\}c^{2}.$$

For ${}^{4}\text{He}$, we have $Z = N = 2$; thus,

Equation:

$$\text{BE} = \{[2m({}^{1}\text{H}) + 2m_{n}] - m({}^{4}\text{He})\}c^{2}.$$

Appendix A gives these masses as $m({}^{4}\text{He}) = 4.002602\ \text{u}$, $m({}^{1}\text{H}) = 1.007825\ \text{u}$, and $m_{n} = 1.008665\ \text{u}$. Thus,

Equation:

$$\text{BE} = (0.030378\ \text{u})c^{2}.$$

Noting that $1\ \text{u} = 931.5\ \text{MeV}/c^{2}$, we find

Equation:

$$\text{BE} = (0.030378)(931.5\ \text{MeV}/c^{2})c^{2} = 28.3\ \text{MeV}.$$

Since $A = 4$, we see that BE/A is this number divided by 4, or

Equation:

$$\text{BE}/A = 7.07\text{ MeV/nucleon.}$$

Discussion

This is a large binding energy per nucleon compared with those for other low-mass nuclei, which have $\text{BE}/A \approx 3\text{ MeV/nucleon}$. This indicates that ^4He is tightly bound compared with its neighbors on the chart of the nuclides. You can see the spike representing this value of BE/A for ^4He on the graph in [link]. This is why ^4He is stable. Since ^4He is tightly bound, it has less mass than other $A = 4$ nuclei and, therefore, cannot spontaneously decay into them. The large binding energy also helps to explain why some nuclei undergo α decay. Smaller mass in the decay products can mean energy release, and such decays can be spontaneous. Further, it can happen that two protons and two neutrons in a nucleus can randomly find themselves together, experience the exceptionally large nuclear force that binds this combination, and act as a ^4He unit within the nucleus, at least for a while. In some cases, the ^4He escapes, and α decay has then taken place.

There is more to be learned from nuclear binding energies. The general trend in BE/A is fundamental to energy production in stars, and to fusion and fission energy sources on Earth, for example. This is one of the applications of nuclear physics covered in Medical Applications of Nuclear Physics. The abundance of elements on Earth, in stars, and in the universe as a whole is related to the binding energy of nuclei and has implications for the continued expansion of the universe.

Problem-Solving Strategies

For Reaction And Binding Energies and Activity Calculations in Nuclear Physics

1. *Identify exactly what needs to be determined in the problem (identify the unknowns)*. This will allow you to decide whether the energy of a decay or nuclear reaction is involved, for example, or whether the problem is primarily concerned with activity (rate of decay).

2. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).*
3. *For reaction and binding-energy problems, we use atomic rather than nuclear masses.* Since the masses of neutral atoms are used, you must count the number of electrons involved. If these do not balance (such as in β^{+} decay), then an energy adjustment of 0.511 MeV per electron must be made. Also note that atomic masses may not be given in a problem; they can be found in tables.
4. *For problems involving activity, the relationship of activity to half-life, and the number of nuclei given in the equation* $R=\frac{0.693N}{t_{1/2}}$ *can be very useful.* Owing to the fact that number of nuclei is involved, you will also need to be familiar with moles and Avogadro's number.
5. *Perform the desired calculation; keep careful track of plus and minus signs as well as powers of 10.*
6. *Check the answer to see if it is reasonable: Does it make sense?* Compare your results with worked examples and other information in the text. (Heeding the advice in Step 5 will also help you to be certain of your result.) You must understand the problem conceptually to be able to determine whether the numerical result is reasonable.

Note:
PhET Explorations: Nuclear Fission
Start a chain reaction, or introduce non-radioactive isotopes to prevent one. Control energy production in a nuclear reactor!

https://archive.cnx.org/specials/01caf0d0-116f-11e6-b891-abfdaa77b03b/nuclear-fission/#sim-one-nucleus

Section Summary

- The binding energy (BE) of a nucleus is the energy needed to separate it into individual protons and neutrons. In terms of atomic masses,
  Equation:

$$\mathrm{BE} = \{[\mathrm{Zm}(^1\mathrm{H}) + \mathrm{Nm}_n] - m(^A\mathrm{X})\}c^2,$$

where $m\left(^1\mathrm{H}\right)$ is the mass of a hydrogen atom, $m\left(^A\mathrm{X}\right)$ is the atomic mass of the nuclide, and m_n is the mass of a neutron. Patterns in the binding energy per nucleon, BE/A, reveal details of the nuclear force. The larger the BE/A, the more stable the nucleus.

Conceptual Questions

Exercise:

Problem:

Why is the number of neutrons greater than the number of protons in stable nuclei having A greater than about 40, and why is this effect more pronounced for the heaviest nuclei?

Problems & Exercises

Exercise:

Problem:

^{2}H is a loosely bound isotope of hydrogen. Called deuterium or heavy hydrogen, it is stable but relatively rare—it is 0.015% of natural hydrogen. Note that deuterium has $Z = N$, which should tend to make it more tightly bound, but both are odd numbers. Calculate BE/A, the binding energy per nucleon, for ^{2}H and compare it with the approximate value obtained from the graph in [link].

Solution:

1.112 MeV, consistent with graph

Exercise:

Problem:

^{56}Fe is among the most tightly bound of all nuclides. It is more than 90% of natural iron. Note that ^{56}Fe has even numbers of both protons and neutrons. Calculate BE/A, the binding energy per nucleon, for ^{56}Fe and compare it with the approximate value obtained from the graph in [link].

Exercise:

Problem:

^{209}Bi is the heaviest stable nuclide, and its BE/A is low compared with medium-mass nuclides. Calculate BE/A, the binding energy per nucleon, for ^{209}Bi and compare it with the approximate value obtained from the graph in [link].

Solution:

7.848 MeV, consistent with graph

Exercise:

Problem:

(a) Calculate BE/A for ^{235}U, the rarer of the two most common uranium isotopes. (b) Calculate BE/A for ^{238}U. (Most of uranium is ^{238}U.) Note that ^{238}U has even numbers of both protons and neutrons. Is the BE/A of ^{238}U significantly different from that of ^{235}U ?

Exercise:

Problem:

(a) Calculate BE/A for ^{12}C. Stable and relatively tightly bound, this nuclide is most of natural carbon. (b) Calculate BE/A for ^{14}C. Is the difference in BE/A between ^{12}C and ^{14}C significant? One is stable and common, and the other is unstable and rare.

Solution:

(a) 7.680 MeV, consistent with graph

(b) 7.520 MeV, consistent with graph. Not significantly different from value for ^{12}C, but sufficiently lower to allow decay into another nuclide that is more tightly bound.

Exercise:

Problem:

The fact that BE/A is greatest for A near 60 implies that the range of the nuclear force is about the diameter of such nuclides. (a) Calculate the diameter of an $A = 60$ nucleus. (b) Compare BE/A for ^{58}Ni and ^{90}Sr. The first is one of the most tightly bound nuclides, while the second is larger and less tightly bound.

Exercise:

Problem:

The purpose of this problem is to show in three ways that the binding energy of the electron in a hydrogen atom is negligible compared with the masses of the proton and electron. (a) Calculate the mass equivalent in u of the 13.6-eV binding energy of an electron in a hydrogen atom, and compare this with the mass of the hydrogen atom obtained from Appendix A. (b) Subtract the mass of the proton given in [link] from the mass of the hydrogen atom given in Appendix A. You will find the difference is equal to the electron’s mass to three digits, implying the binding energy is small in comparison. (c) Take the ratio of the binding energy of the electron (13.6 eV) to the energy equivalent of the electron’s mass (0.511 MeV). (d) Discuss how your answers confirm the stated purpose of this problem.

Solution:

(a) 1.46×10^{-8} u vs. 1.007825 u for ^{1}H

(b) 0.000549 u

(c) 2.66×10^{-5}

Exercise:

Problem: **Unreasonable Results**

A particle physicist discovers a neutral particle with a mass of 2.02733 u that he assumes is two neutrons bound together. (a) Find the binding energy. (b) What is unreasonable about this result? (c) What assumptions are unreasonable or inconsistent?

Solution:

(a) –9.315 MeV

(b) The negative binding energy implies an unbound system.

(c) This assumption that it is two bound neutrons is incorrect.

Glossary

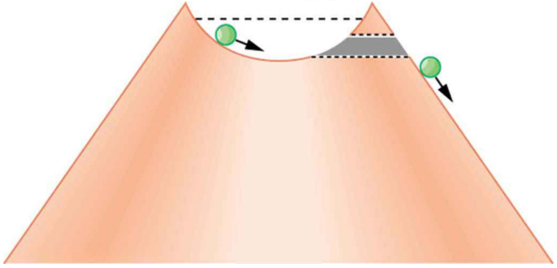
binding energy
    the energy needed to separate nucleus into individual protons and neutrons

binding energy per nucleon
    the binding energy calculated per nucleon; it reveals the details of the nuclear force—larger the BE/A, the more stable the nucleus

Tunneling

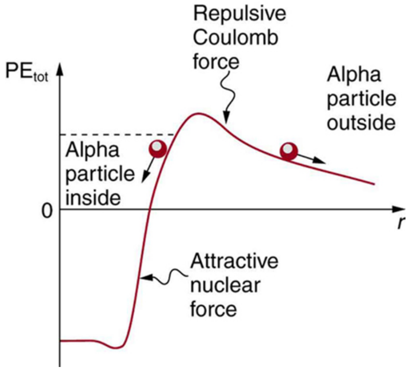
- Define and discuss tunneling.
- Define potential barrier.
- Explain quantum tunneling.

Protons and neutrons are *bound* inside nuclei, that means energy must be supplied to break them away. The situation is analogous to a marble in a bowl that can roll around but lacks the energy to get over the rim. It is bound inside the bowl (see [link]). If the marble could get over the rim, it would gain kinetic energy by rolling down outside. However classically, if the marble does not have enough kinetic energy to get over the rim, it remains forever trapped in its well.



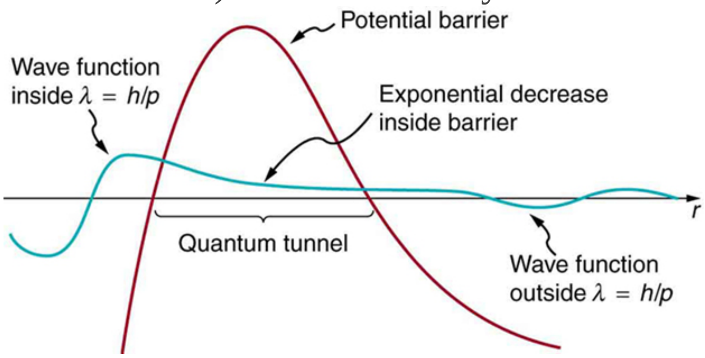
The marble in this semicircular bowl at the top of a volcano has enough kinetic energy to get to the altitude of the dashed line, but not enough to get over the rim, so that it is trapped forever. If it could find a tunnel through the barrier, it would escape, roll downhill, and gain kinetic energy.

In a nucleus, the attractive nuclear potential is analogous to the bowl at the top of a volcano (where the “volcano” refers only to the shape). Protons and neutrons have kinetic energy, but it is about 8 MeV less than that needed to get out (see [link]). That is, they are bound by an average of 8 MeV per nucleon. The slope of the hill outside the bowl is analogous to the repulsive Coulomb potential for a nucleus, such as for an α particle outside a positive nucleus. In α decay, two protons and two neutrons spontaneously break away as a ^4He unit. Yet the protons and neutrons do not have enough kinetic energy to get over the rim. So how does the α particle get out?



Nucleons within an atomic nucleus are bound or trapped by the attractive nuclear force, as shown in this simplified potential energy curve. An α particle outside the range of the nuclear force feels the repulsive Coulomb force. The α particle inside the nucleus does not have enough kinetic energy to get over the rim, yet it does manage to get out by quantum mechanical tunneling.

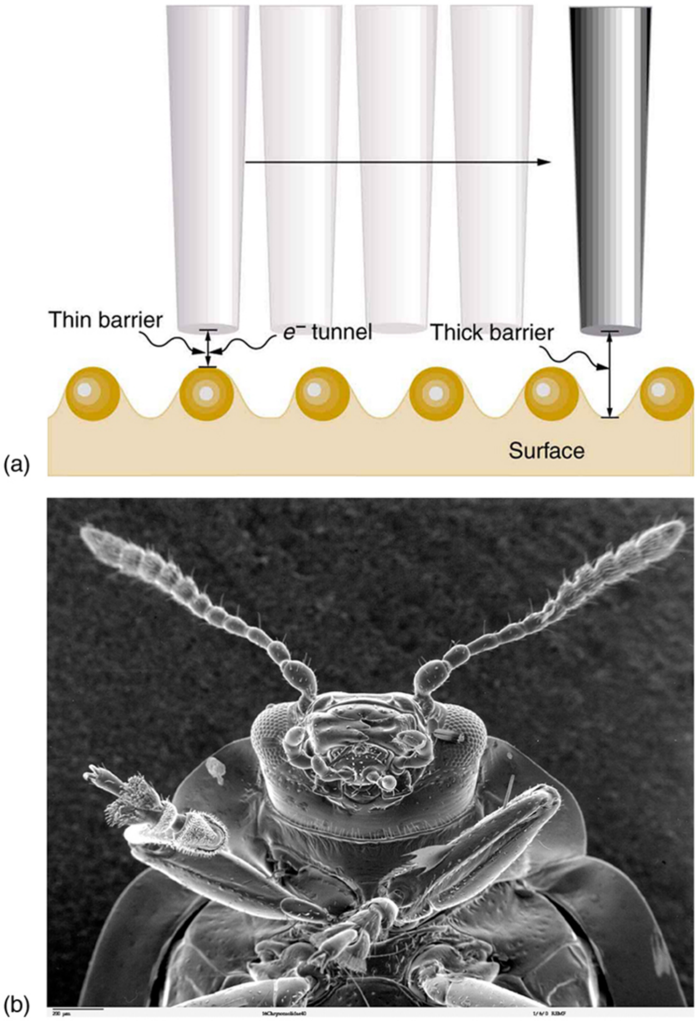
The answer was supplied in 1928 by the Russian physicist George Gamow (1904–1968). The α particle tunnels through a region of space it is forbidden to be in, and it comes out of the side of the nucleus. Like an electron making a transition between orbits around an atom, it travels from one point to another without ever having been in between. [link] indicates how this works. The wave function of a quantum mechanical particle varies smoothly, going from within an atomic nucleus (on one side of a potential energy barrier) to outside the nucleus (on the other side of the potential energy barrier). Inside the barrier, the wave function does not become zero but decreases exponentially, and we do not observe the particle inside the barrier. The probability of finding a particle is related to the square of its wave function, and so there is a small probability of finding the particle outside the barrier, which implies that the particle can tunnel through the barrier. This process is called **barrier penetration** or **quantum mechanical tunneling**. This concept was developed in theory by J. Robert Oppenheimer (who led the development of the first nuclear bombs during World War II) and was used by Gamow and others to describe α decay.


Potential barrier
Wave function inside λ = h/p
Exponential decrease inside barrier
r
Quantum tunnel
Wave function outside λ = h/p

The wave function representing a quantum mechanical particle must vary smoothly, going from within the nucleus (to the left of the barrier) to outside the nucleus (to the right of the barrier). Inside the barrier, the wave function does not abruptly become zero; rather, it decreases exponentially. Outside the barrier, the wave function is small but finite, and there it smoothly becomes sinusoidal. Owing to the fact that there is a small probability of finding the particle outside the barrier, the particle can tunnel through the barrier.

Good ideas explain more than one thing. In addition to qualitatively explaining how the four nucleons in an α particle can get out of the nucleus, the detailed theory also explains quantitatively the half-life of various nuclei that undergo α decay. This description is what Gamow and others devised, and it works for α decay half-lives that vary by 17 orders of magnitude. Experiments have shown that the more energetic the α decay of a particular nuclide is, the shorter is its half-life. **Tunneling** explains this in the following manner: For the decay to be more energetic, the nucleons must have more energy in the nucleus and should be able to ascend a little closer to the rim. The barrier is therefore not as thick for more energetic decay, and the exponential decrease of the wave function inside the barrier is not as great. Thus the probability of finding the particle outside the barrier is greater, and the half-life is shorter.

Tunneling as an effect also occurs in quantum mechanical systems other than nuclei. Electrons trapped in solids can tunnel from one object to another if the barrier between the objects is thin enough. The process is the same in principle as described for α decay. It is far more likely for a thin barrier than a thick one. Scanning tunneling electron microscopes function on this principle. The current of electrons that travels between a probe and a sample tunnels through a barrier and is very sensitive to its thickness, allowing detection of individual atoms as shown in [link].



(a) A scanning tunneling electron microscope can detect extremely small variations in dimensions, such as individual atoms. Electrons tunnel quantum mechanically between the probe and the sample. The probability of tunneling is extremely sensitive to barrier thickness, so that the electron current is a sensitive indicator of surface features. (b) Head and mouthparts of *Coleoptera Chrysomelidea* as seen through an electron microscope (credit: Louisa Howard, Dartmouth College)

Note:

PhET Explorations: Quantum Tunneling and Wave Packets

Watch quantum "particles" tunnel through barriers. Explore the properties of the wave functions that describe these particles.

Quantum Tunneling and Wave Packets

Section Summary

- Tunneling is a quantum mechanical process of potential energy barrier penetration. The concept was first applied to explain α decay, but tunneling is found to occur in other quantum mechanical systems.

Conceptual Questions

Exercise:

Problem:

A physics student caught breaking conservation laws is imprisoned. She leans against the cell wall hoping to tunnel out quantum mechanically. Explain why her chances are negligible. (This is so in any classical situation.)

Exercise:

Problem:

When a nucleus α decays, does the α particle move continuously from inside the nucleus to outside? That is, does it travel each point along an imaginary line from inside to out? Explain.

Problems-Exercises

Exercise:

Problem:

Derive an approximate relationship between the energy of α decay and half-life using the following data. It may be useful to graph the log of $t_{1/2}$ against E_α to find some straight-line relationship.

Nuclide	E_α (MeV)	$t_{1/2}$
^{216}Ra	9.5	0.18 μs
^{194}Po	7.0	0.7 s
^{240}Cm	6.4	27 d
^{226}Ra	4.91	1600 y
^{232}Th	4.1	1.4×10^{10} y

Energy and Half-Life for α Decay

Exercise:

Problem: **Integrated Concepts**

A 2.00-T magnetic field is applied perpendicular to the path of charged particles in a bubble chamber. What is the radius of curvature of the path of a 10 MeV proton in this field? Neglect any slowing along its path.

Solution:

22.8 cm

Exercise:

Problem:

(a) Write the decay equation for the α decay of ^{235}U. (b) What energy is released in this decay? The mass of the daughter nuclide is 231.036298 u. (c) Assuming the residual nucleus is formed in its ground state, how much energy goes to the α particle?

Solution:

(a) $^{235}_{92}\text{U}_{143} \rightarrow {}^{231}_{90}\text{Th}_{141} + {}^{4}_{2}\text{He}_{2}$

(b) 4.679 MeV

(c) 4.599 MeV

Exercise:

Problem: **Unreasonable Results**

The relatively scarce naturally occurring calcium isotope ^{48}Ca has a half-life of about 2×10^{16} y. (a) A small sample of this isotope is labeled as having an activity of 1.0 Ci. What is the mass of the ^{48}Ca in the sample? (b) What is unreasonable about this result? (c) What assumption is responsible?

Exercise:

Problem: **Unreasonable Results**

A physicist scatters γ rays from a substance and sees evidence of a nucleus 7.5×10^{-13} m in radius. (a) Find the atomic mass of such a nucleus. (b) What is unreasonable about this result? (c) What is unreasonable about the assumption?

Solution:

a) 2.4×10^{8} u

(b) The greatest known atomic masses are about 260. This result found in (a) is extremely large.

(c) The assumed radius is much too large to be reasonable.

Exercise:

Problem: **Unreasonable Results**

A frazzled theoretical physicist reckons that all conservation laws are obeyed in the decay of a proton into a neutron, positron, and neutrino (as in β^{+} decay of a nucleus) and sends a paper to a journal to announce the reaction as a possible end of the universe due to the spontaneous decay of protons. (a) What energy is released in this decay? (b) What is unreasonable about this result? (c) What assumption is responsible?

Solution:

(a) -1.805 MeV

(b) Negative energy implies energy input is necessary and the reaction cannot be spontaneous.

(c) Although all conversation laws are obeyed, energy must be supplied, so the assumption of spontaneous decay is incorrect.

Exercise:

Problem: Construct Your Own Problem

Consider the decay of radioactive substances in the Earth’s interior. The energy emitted is converted to thermal energy that reaches the earth’s surface and is radiated away into cold dark space. Construct a problem in which you estimate the activity in a cubic meter of earth rock? And then calculate the power generated. Calculate how much power must cross each square meter of the Earth’s surface if the power is dissipated at the same rate as it is generated. Among the things to consider are the activity per cubic meter, the energy per decay, and the size of the Earth.

Glossary

barrier penetration

quantum mechanical effect whereby a particle has a nonzero probability to cross through a potential energy barrier despite not having sufficient energy to pass over the barrier; also called quantum mechanical tunneling

quantum mechanical tunneling

quantum mechanical effect whereby a particle has a nonzero probability to cross through a potential energy barrier despite not having sufficient energy to pass over the barrier; also called barrier penetration

tunneling

a quantum mechanical process of potential energy barrier penetration

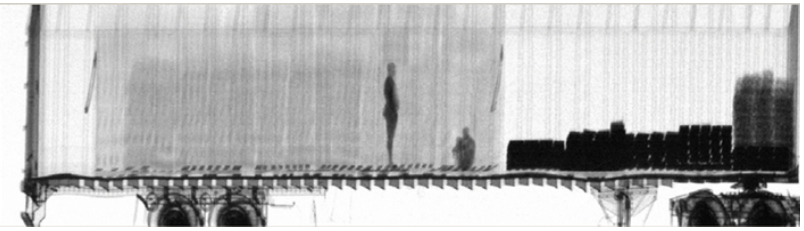
Introduction to Applications of Nuclear Physics
class="introduction"

- Provide examples of various nuclear physics applications.

Tori Randall, Ph.D., curator for the Department of Physical Anthropology at the San Diego Museum of Man, prepares a 550-year-old Peruvian child mummy for a CT scan at Naval Medical Center San Diego. (credit: U.S. Navy photo by Mass Communication Specialist 3rd Class Samantha A. Lewis)

Applications of nuclear physics have become an integral part of modern life. From the bone scan that detects a cancer to the radioiodine treatment that cures another, nuclear radiation has diagnostic and therapeutic effects on medicine. From the fission power reactor to the hope of controlled fusion, nuclear energy is now commonplace and is a part of our plans for the future. Yet, the destructive potential of nuclear weapons haunts us, as does the possibility of nuclear reactor accidents. Certainly, several applications of nuclear physics escape our view, as seen in [link]. Not only has nuclear physics revealed secrets of nature, it has an inevitable impact based on its applications, as they are intertwined with human values. Because of its potential for alleviation of suffering, and its power as an ultimate destructor of life, nuclear physics is often viewed with ambivalence. But it provides perhaps the best example that applications can be good or evil, while knowledge itself is neither.

Customs officers inspect vehicles using neutron irradiation. Cars and trucks pass through portable x-ray machines that reveal their contents. (credit: Gerald L. Nino, CBP, U.S. Dept. of Homeland Security)

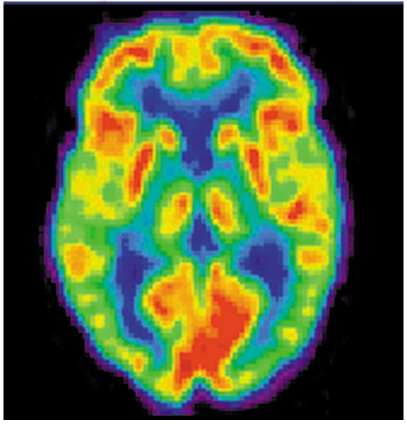


This image shows two stowaways caught illegally entering the United States from Canada. (credit: U.S. Customs and Border Protection)

Medical Imaging and Diagnostics

- Explain the working principle behind an anger camera.
- Describe the SPECT and PET imaging techniques.

A host of medical imaging techniques employ nuclear radiation. What makes nuclear radiation so useful? First, γ radiation can easily penetrate tissue; hence, it is a useful probe to monitor conditions inside the body. Second, nuclear radiation depends on the nuclide and not on the chemical compound it is in, so that a radioactive nuclide can be put into a compound designed for specific purposes. The compound is said to be **tagged**. A tagged compound used for medical purposes is called a **radiopharmaceutical**. Radiation detectors external to the body can determine the location and concentration of a radiopharmaceutical to yield medically useful information. For example, certain drugs are concentrated in inflamed regions of the body, and this information can aid diagnosis and treatment as seen in [link]. Another application utilizes a radiopharmaceutical which the body sends to bone cells, particularly those that are most active, to detect cancerous tumors or healing points. Images can then be produced of such bone scans. Radioisotopes are also used to determine the functioning of body organs, such as blood flow, heart muscle activity, and iodine uptake in the thyroid gland.



A radiopharmaceutical is used to produce this brain image of a patient with

Alzheimer’s disease. Certain features are computer enhanced. (credit: National Institutes of Health)

Medical Application

[link] lists certain medical diagnostic uses of radiopharmaceuticals, including isotopes and activities that are typically administered. Many organs can be imaged with a variety of nuclear isotopes replacing a stable element by a radioactive isotope. One common diagnostic employs iodine to image the thyroid, since iodine is concentrated in that organ. The most active thyroid cells, including cancerous cells, concentrate the most iodine and, therefore, emit the most radiation. Conversely, hypothyroidism is indicated by lack of iodine uptake. Note that there is more than one isotope that can be used for several types of scans. Another common nuclear diagnostic is the thallium scan for the cardiovascular system, particularly used to evaluate blockages in the coronary arteries and examine heart activity. The salt TlCl can be used, because it acts like NaCl and follows the blood. Gallium-67 accumulates where there is rapid cell growth, such as in tumors and sites of infection. Hence, it is useful in cancer imaging. Usually, the patient receives the injection one day and has a whole body scan 3 or 4 days later because it can take several days for the gallium to build up.

Procedure, isotope	**Typical activity (mCi), where** $1\text{ mCi}=3.7\times 10^{7}\text{ Bq}$

Procedure, isotope	**Typical activity (mCi), where** $1\ \text{mCi} = 3.7 \times 10^7\ \text{Bq}$
Brain scan	
$^{99\text{m}}\text{Tc}$	7.5
$^{113\text{m}}\text{In}$	7.5
^{11}C (PET)	20
^{13}N (PET)	20
^{15}O (PET)	50
^{18}F (PET)	10
Lung scan	
$^{99\text{m}}\text{Tc}$	2

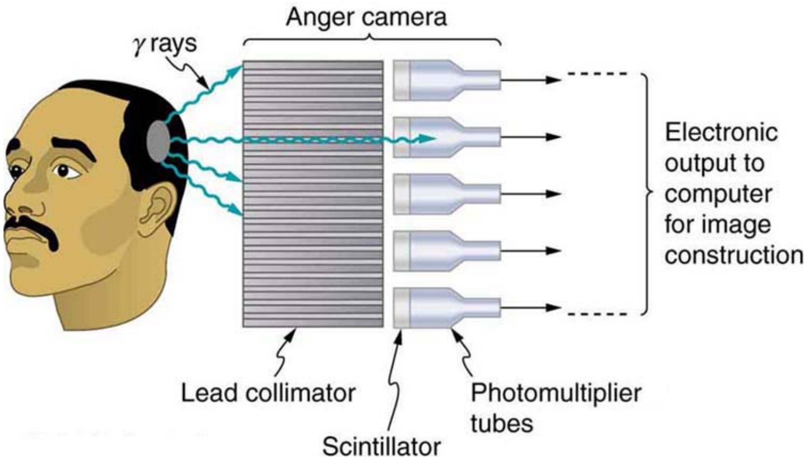
Procedure, isotope	**Typical activity (mCi), where** $1 \text{ mCi} = 3.7 \times 10^7 \text{ Bq}$
^{133}Xe	7.5
Cardiovascular blood pool	
^{131}I	0.2
$^{99\text{m}}\text{Tc}$	2
Cardiovascular arterial flow	
^{201}Tl	3
^{24}Na	7.5
Thyroid scan	
^{131}I	0.05
^{123}I	0.07

Procedure, isotope	**Typical activity (mCi), where** $1 \text{ mCi} = 3.7 \times 10^7 \text{ Bq}$
Liver scan	
^{198}Au (colloid)	0.1
$^{99\text{m}}\text{Tc}$ (colloid)	2
Bone scan	
^{85}Sr	0.1
$^{99\text{m}}\text{Tc}$	10
Kidney scan	
^{197}Hg	0.1
$^{99\text{m}}\text{Tc}$	1.5

Diagnostic Uses of Radiopharmaceuticals

Note that [link] lists many diagnostic uses for $^{99\text{m}}\text{Tc}$, where “m” stands for a metastable state of the technetium nucleus. Perhaps 80 percent of all radiopharmaceutical procedures employ $^{99\text{m}}\text{Tc}$ because of its many advantages. One is that the decay of its metastable state produces a single, easily identified 0.142-MeV γ ray. Additionally, the radiation dose to the patient is limited by the short 6.0-h half-life of $^{99\text{m}}\text{Tc}$. And, although its half-life is short, it is easily and continuously produced on site. The basic process for production is neutron activation of molybdenum, which quickly β decays into $^{99\text{m}}\text{Tc}$. Technetium-99m can be attached to many compounds to allow the imaging of the skeleton, heart, lungs, kidneys, etc.

[link] shows one of the simpler methods of imaging the concentration of nuclear activity, employing a device called an **Anger camera** or **gamma camera**. A piece of lead with holes bored through it collimates γ rays emerging from the patient, allowing detectors to receive γ rays from specific directions only. The computer analysis of detector signals produces an image. One of the disadvantages of this detection method is that there is no depth information (i.e., it provides a two-dimensional view of the tumor as opposed to a three-dimensional view), because radiation from any location under that detector produces a signal.

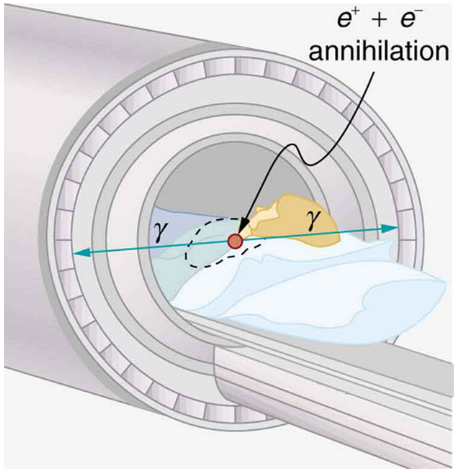


An Anger or gamma camera consists of a lead collimator and an array of detectors. Gamma rays produce light flashes in the scintillators. The light output is converted to an electrical signal by the photomultipliers. A computer constructs an image from the detector output.

Imaging techniques much like those in x-ray computed tomography (CT) scans use nuclear activity in patients to form three-dimensional images. [link] shows a patient in a circular array of detectors that may be stationary or rotated, with detector output used by a computer to construct a detailed image. This technique is called **single-photon-emission computed tomography(SPECT)** or sometimes simply SPET. The spatial resolution of this technique is poor, about 1 cm, but the contrast (i.e. the difference in visual properties that makes an object distinguishable from other objects and the background) is good.

SPECT uses a geometry similar to a CT scanner to form an image of the concentration of a radiopharmaceutical compound. (credit: Woldo, Wikimedia Commons)

Images produced by β^+ emitters have become important in recent years. When the emitted positron (β^+) encounters an electron, mutual annihilation occurs, producing two γ rays. These γ rays have identical 0.511-MeV energies (the energy comes from the destruction of an electron or positron mass) and they move directly away from one another, allowing detectors to determine their point of origin accurately, as shown in [link]. The system is called **positron emission tomography (PET)**. It requires detectors on opposite sides to simultaneously (i.e., at the same time) detect photons of 0.511-MeV energy and utilizes computer imaging techniques similar to those in SPECT and CT scans. Examples of β^+ -emitting isotopes used in PET are ^{11}C, ^{13}N, ^{15}O, and ^{18}F, as seen in [link]. This list includes C, N, and O, and so they have the advantage of being able to function as tags for natural body compounds. Its resolution of 0.5 cm is better than that of SPECT; the accuracy and sensitivity of PET scans make them useful for examining the brain's anatomy and function. The brain's use of oxygen and water can be monitored with ^{15}O. PET is used extensively for diagnosing brain disorders. It can note decreased metabolism in certain regions prior to a confirmation of Alzheimer's disease. PET can locate regions in the brain that become active when a person carries out specific activities, such as speaking, closing their eyes, and so on.



A PET system takes

advantage of the two identical γ-ray photons produced by positron-electron annihilation. These γ rays are emitted in opposite directions, so that the line along which each pair is emitted is determined. Various events detected by several pairs of detectors are then analyzed by the computer to form an accurate image.

Note:

PhET Explorations: Simplified MRI

Is it a tumor? Magnetic Resonance Imaging (MRI) can tell. Your head is full of tiny radio transmitters (the nuclear spins of the hydrogen nuclei of your water molecules). In an MRI unit, these little radios can be made to broadcast their positions, giving a detailed picture of the inside of your head.

Simplified MRI

Section Summary

- Radiopharmaceuticals are compounds that are used for medical imaging and therapeutics.
- The process of attaching a radioactive substance is called tagging.
- [link] lists certain diagnostic uses of radiopharmaceuticals including the isotope and activity typically used in diagnostics.
- One common imaging device is the Anger camera, which consists of a lead collimator, radiation detectors, and an analysis computer.
- Tomography performed with γ-emitting radiopharmaceuticals is called SPECT and has the advantages of x-ray CT scans coupled with organ- and function-specific drugs.
- PET is a similar technique that uses β^+ emitters and detects the two annihilation γ rays, which aid to localize the source.

Conceptual Questions

Exercise:

Problem:

In terms of radiation dose, what is the major difference between medical diagnostic uses of radiation and medical therapeutic uses?

Exercise:

Problem:

One of the methods used to limit radiation dose to the patient in medical imaging is to employ isotopes with short half-lives. How would this limit the dose?

Problems & Exercises

Exercise:

Problem:

A neutron generator uses an α source, such as radium, to bombard beryllium, inducing the reaction ${}^{4}\text{He}+{}^{9}\text{Be}\rightarrow{}^{12}\text{C}+n$. Such neutron sources are called RaBe sources, or PuBe sources if they use plutonium to get the α s. Calculate the energy output of the reaction in MeV.

Solution:

5.701 MeV

Exercise:

Problem:

Neutrons from a source (perhaps the one discussed in the preceding problem) bombard natural molybdenum, which is 24 percent ${}^{98}\text{Mo}$. What is the energy output of the reaction ${}^{98}\text{Mo}+n\rightarrow{}^{99}\text{Mo}+\gamma$? The mass of ${}^{98}\text{Mo}$ is given in Appendix A: Atomic Masses, and that of ${}^{99}\text{Mo}$ is 98.907711 u.

Exercise:

Problem:

The purpose of producing ${}^{99}\text{Mo}$ (usually by neutron activation of natural molybdenum, as in the preceding problem) is to produce ${}^{99\text{m}}\text{Tc}$. Using the rules, verify that the β^{-} decay of ${}^{99}\text{Mo}$ produces ${}^{99\text{m}}\text{Tc}$. (Most ${}^{99\text{m}}\text{Tc}$ nuclei produced in this decay are left in a metastable excited state denoted ${}^{99\text{m}}\text{Tc}$.)

Solution:

${}^{99}_{42}\text{Mo}_{57}\rightarrow{}^{99}_{43}\text{Tc}_{56}+\beta^{-}+v_{e}$

Exercise:

Problem:

(a) Two annihilation γ rays in a PET scan originate at the same point and travel to detectors on either side of the patient. If the point of origin is 9.00 cm closer to one of the detectors, what is the difference in arrival times of the photons? (This could be used to give position information, but the time difference is small enough to make it difficult.)

(b) How accurately would you need to be able to measure arrival time differences to get a position resolution of 1.00 mm?

Exercise:

Problem:

[link] indicates that 7.50 mCi of ^{99m}Tc is used in a brain scan. What is the mass of technetium?

Solution:

1.43×10^{-9} g

Exercise:

Problem:

The activities of ^{131}I and ^{123}I used in thyroid scans are given in [link] to be 50 and $70\ \mu\text{Ci}$, respectively. Find and compare the masses of ^{131}I and ^{123}I in such scans, given their respective half-lives are 8.04 d and 13.2 h. The masses are so small that the radioiodine is usually mixed with stable iodine as a carrier to ensure normal chemistry and distribution in the body.

Exercise:

Problem:

(a) Neutron activation of sodium, which is 100% ^{23}Na, produces ^{24}Na, which is used in some heart scans, as seen in [link]. The equation for the reaction is $^{23}\text{Na} + n \rightarrow {}^{24}\text{Na} + \gamma$. Find its energy output, given the mass of ^{24}Na is 23.990962 u.

(b) What mass of ^{24}Na produces the needed 5.0-mCi activity, given its half-life is 15.0 h?

Solution:

(a) 6.958 MeV

(b) 5.7×10^{-10} g

Glossary

Anger camera
: a common medical imaging device that uses a scintillator connected to a series of photomultipliers

gamma camera
: another name for an Anger camera

positron emission tomography (PET)
: tomography technique that uses β^{+} emitters and detects the two annihilation γ rays, aiding in source localization

radiopharmaceutical
: compound used for medical imaging

single-photon-emission computed tomography (SPECT)
: tomography performed with γ-emitting radiopharmaceuticals

tagged
: process of attaching a radioactive substance to a chemical compound

Biological Effects of Ionizing Radiation

- Define various units of radiation.
- Describe RBE.

We hear many seemingly contradictory things about the biological effects of ionizing radiation. It can cause cancer, burns, and hair loss, yet it is used to treat and even cure cancer. How do we understand these effects? Once again, there is an underlying simplicity in nature, even in complicated biological organisms. All the effects of ionizing radiation on biological tissue can be understood by knowing that **ionizing radiation affects molecules within cells, particularly DNA molecules.**

Let us take a brief look at molecules within cells and how cells operate. Cells have long, double-helical DNA molecules containing chemical codes called genetic codes that govern the function and processes undertaken by the cell. It is for unraveling the double-helical structure of DNA that James Watson, Francis Crick, and Maurice Wilkins received the Nobel Prize. Damage to DNA consists of breaks in chemical bonds or other changes in the structural features of the DNA chain, leading to changes in the genetic code. In human cells, we can have as many as a million individual instances of damage to DNA per cell per day. It is remarkable that DNA contains codes that check whether the DNA is damaged or can repair itself. It is like an auto check and repair mechanism. This repair ability of DNA is vital for maintaining the integrity of the genetic code and for the normal functioning of the entire organism. It should be constantly active and needs to respond rapidly. The rate of DNA repair depends on various factors such as the cell type and age of the cell. A cell with a damaged ability to repair DNA, which could have been induced by ionizing radiation, can do one of the following:

- The cell can go into an irreversible state of dormancy, known as senescence.
- The cell can commit suicide, known as programmed cell death.
- The cell can go into unregulated cell division leading to tumors and cancers.

Since ionizing radiation damages the DNA, which is critical in cell reproduction, it has its greatest effect on cells that rapidly reproduce, including most types of cancer. Thus, cancer cells are more sensitive to radiation than normal cells and can be killed by it easily. Cancer is characterized by a malfunction of cell reproduction, and can also be caused by ionizing radiation. Without contradiction, ionizing radiation can be both a cure and a cause.

To discuss quantitatively the biological effects of ionizing radiation, we need a radiation dose unit that is directly related to those effects. All effects of radiation are assumed to be directly proportional to the amount of ionization produced in the biological organism. The amount of ionization is in turn proportional to the amount of deposited energy. Therefore, we define a **radiation dose unit** called the **rad**, as $1/100$ of a joule of ionizing energy deposited per kilogram of tissue, which is

Equation:

$$1\text{ rad}=0.01\text{ J/kg}.$$

For example, if a 50.0-kg person is exposed to ionizing radiation over her entire body and she absorbs 1.00 J, then her whole-body radiation dose is

Equation:

$$(1.00\text{ J})/(50.0\text{ kg})=0.0200\text{ J/kg}=2.00\text{ rad}.$$

If the same 1.00 J of ionizing energy were absorbed in her 2.00-kg forearm alone, then the dose to the forearm would be

Equation:

$$(1.00\text{ J})/(2.00\text{ kg})=0.500\text{ J/kg}=50.0\text{ rad},$$

and the unaffected tissue would have a zero rad dose. While calculating radiation doses, you divide the energy absorbed by the mass of affected tissue. You must specify the affected region, such as the whole body or forearm in addition to giving the numerical dose in rads. The SI unit for radiation dose is the **gray (Gy)**, which is defined to be

Equation:

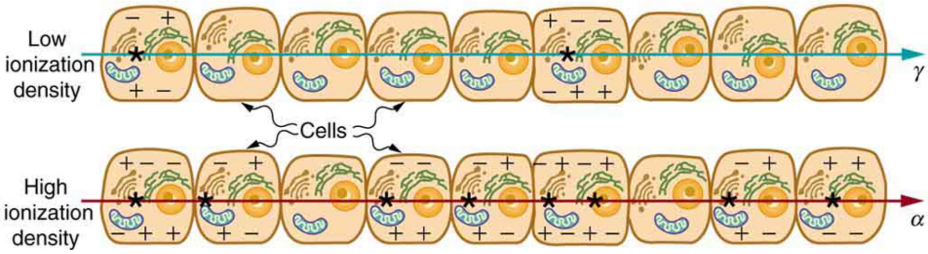
$$1\text{ Gy}=1\text{ J/kg}=100\text{ rad}.$$

However, the rad is still commonly used. Although the energy per kilogram in 1 rad is small, it has significant effects since the energy causes ionization. The energy needed for a single ionization is a few eV, or less than 10^{-18} J. Thus, 0.01 J of ionizing energy can create a huge number of ion pairs and have an effect at the cellular level.

The effects of ionizing radiation may be directly proportional to the dose in rads, but they also depend on the type of radiation and the type of tissue. That is, for a given dose in rads, the effects depend on whether the radiation is α, β, γ, x-ray, or some other type of ionizing radiation. In the earlier discussion of the range of ionizing radiation, it was noted that energy is deposited in a series of ionizations and not in a single interaction. Each ion pair or ionization requires a certain amount of energy, so that the number of ion pairs is directly proportional to the amount of the deposited ionizing energy. But, if the range of the radiation is small, as it is for α s, then the ionization and the damage created is more concentrated and harder for the organism to repair, as seen in [link]. Concentrated damage is more difficult for biological organisms to repair than damage that is spread out, so short-range particles have greater biological effects. The **relative biological effectiveness** (RBE) or **quality factor** (QF) is given in [link] for several types of ionizing radiation—the effect of the radiation is directly proportional to the RBE. A dose unit more closely related to effects in biological tissue is called the **roentgen equivalent man** or rem and is defined to be the dose in rads multiplied by the relative biological effectiveness.

Equation:

$$\text{rem} = \text{rad} \times \text{RBE}$$



The image shows ionization created in cells by α and γ radiation. Because of its shorter range, the ionization and damage created by α is more concentrated and harder for the organism to repair. Thus, the RBE for α s is greater than the RBE for γ s, even though they create the same amount of ionization at the same energy.

So, if a person had a whole-body dose of 2.00 rad of γ radiation, the dose in rem would be $(2.00\text{ rad})(1) = 2.00\text{ rem whole body}$. If the person had a whole-body dose of 2.00 rad of α radiation, then the dose in rem would be $(2.00\text{ rad})(20) = 40.0\text{ rem whole body}$. The α s would have 20 times the effect on the person than the γ s for the same deposited energy. The SI equivalent of the rem is the **sievert** (Sv), defined to be $\text{Sv} = \text{Gy} \times \text{RBE}$, so that

Equation:

$$1\text{ Sv} = 1\text{ Gy} \times \text{RBE} = 100\text{ rem}.$$

The RBEs given in [link] are approximate, but they yield certain insights. For example, the eyes are more sensitive to radiation, because the cells of the lens do not repair themselves. Neutrons cause more damage than γ rays, although both are neutral and have large ranges, because neutrons often cause secondary radiation when they are captured. Note that the RBEs are 1 for higher-energy β s, γ s, and x-rays, three of the most common types of radiation. For those types of radiation, the numerical values of the dose in rem and rad are identical. For example, 1 rad of γ radiation is also 1 rem. For that reason, rads are still widely quoted rather than rem. [link] summarizes the units that are used for radiation.

Note:

Misconception Alert: Activity vs. Dose

“Activity” refers to the radioactive source while “dose” refers to the amount of energy from the radiation that is deposited in a person or object.

A high level of activity doesn’t mean much if a person is far away from the source. The activity R of a source depends upon the quantity of material (kg) as well as the half-life. A short half-life will produce many more disintegrations per second. Recall that $R = \frac{0.693N}{t_{1/2}}$. Also, the activity decreases exponentially, which is seen in the equation $R = R_0 e^{-\lambda t}$.

Type and energy of radiation	**RBE[footnote]** **Values approximate, difficult to determine.**
X-rays	1
γ rays	1
β rays greater than 32 keV	1
βrays less than 32 keV	1.7
Neutrons, thermal to slow (<20 keV)	2–5
Neutrons, fast (1–10 MeV)	10 (body), 32 (eyes)
Protons (1–10 MeV)	10 (body), 32 (eyes)
α rays from radioactive decay	10–20
Heavy ions from accelerators	10–20

Relative Biological Effectiveness

Quantity	SI unit name	Definition	Former unit	Conversion
Activity	Becquerel (bq)	decay/sec	Curie (Ci)	$1\text{ Bq}=2.7\times 10^{-11}\text{ Ci}$
Absorbed dose	Gray (Gy)	1 J/kg	rad	$\text{Gy}=100\text{ rad}$
Dose Equivalent	Sievert (Sv)	1 J/kg × RBE	rem	$\text{Sv}=100\text{ rem}$

Units for Radiation

The large-scale effects of radiation on humans can be divided into two categories: immediate effects and long-term effects. [link] gives the immediate effects of whole-body exposures received in less than one day. If the radiation exposure is spread out over more time, greater doses are needed to cause the effects listed. This is due to the body's ability to partially repair the damage. Any dose less than 100 mSv (10 rem) is called a **low dose**, 0.1 Sv to 1 Sv (10 to 100 rem) is called a **moderate dose**, and anything greater than 1 Sv (100 rem) is called a **high dose**. There is no known way to determine after the fact if a person has been exposed to less than 10 mSv.

Dose in Sv [footnote] Multiply by 100 to obtain dose in rem.	Effect
0–0.10	No observable effect.
0.1 – 1	Slight to moderate decrease in white blood cell counts.
0.5	Temporary sterility; 0.35 for women, 0.50 for men.
1 – 2	Significant reduction in blood cell counts, brief nausea and vomiting. Rarely fatal.
2 – 5	Nausea, vomiting, hair loss, severe blood damage, hemorrhage, fatalities.

Dose in Sv [footnote] Multiply by 100 to obtain dose in rem.	**Effect**
4.5	LD50/32. Lethal to 50% of the population within 32 days after exposure if not treated.
5 – 20	Worst effects due to malfunction of small intestine and blood systems. Limited survival.
>20	Fatal within hours due to collapse of central nervous system.

Immediate Effects of Radiation (Adults, Whole Body, Single Exposure)

Immediate effects are explained by the effects of radiation on cells and the sensitivity of rapidly reproducing cells to radiation. The first clue that a person has been exposed to radiation is a change in blood count, which is not surprising since blood cells are the most rapidly reproducing cells in the body. At higher doses, nausea and hair loss are observed, which may be due to interference with cell reproduction. Cells in the lining of the digestive system also rapidly reproduce, and their destruction causes nausea. When the growth of hair cells slows, the hair follicles become thin and break off. High doses cause significant cell death in all systems, but the lowest doses that cause fatalities do so by weakening the immune system through the loss of white blood cells.

The two known long-term effects of radiation are cancer and genetic defects. Both are directly attributable to the interference of radiation with cell reproduction. For high doses of radiation, the risk of cancer is reasonably well known from studies of exposed groups. Hiroshima and Nagasaki survivors and a smaller number of people exposed by their occupation, such as radium dial painters, have been fully documented. Chernobyl victims will be studied for many decades, with some data already available. For example, a significant increase in childhood thyroid cancer has been observed. The risk of a radiation-induced cancer for low and moderate doses is generally *assumed* to be proportional to the risk known for high doses. Under this assumption, any dose of radiation, no matter how small, involves a risk to human health. This is called the **linear hypothesis** and it may be prudent, but it *is* controversial. There is some evidence that, unlike the immediate effects of radiation, the long-term effects are cumulative and there is little self-repair. This is analogous to the risk of skin cancer from UV exposure, which is known to be cumulative.

There is a latency period for the onset of radiation-induced cancer of about 2 years for leukemia and 15 years for most other forms. The person is at risk for at least 30 years after the latency period. Omitting many details, the overall risk of a radiation-induced cancer death per year per rem of exposure is about 10 in a million, which can be written as $10/10^6\ \text{rem}\cdot\text{y}$.

If a person receives a dose of 1 rem, his risk each year of dying from radiation-induced cancer is 10 in a million and that risk continues for about 30 years. The lifetime risk is thus 300 in a million, or 0.03 percent. Since about 20 percent of all worldwide deaths are from cancer, the increase due to a 1 rem exposure is impossible to detect demographically. But 100 rem (1 Sv), which was the dose received by the average Hiroshima and Nagasaki survivor, causes a 3 percent risk, which can be observed in the presence of a 20 percent normal or natural incidence rate.

The incidence of genetic defects induced by radiation is about one-third that of cancer deaths, but is much more poorly known. The lifetime risk of a genetic defect due to a 1 rem exposure is about 100 in a million or $3.3/10^6\ \text{rem}\cdot\text{y}$, but the normal incidence is 60,000 in a million. Evidence of such a small increase, tragic as it is, is nearly impossible to obtain. For example, there is no evidence of increased genetic defects among the offspring of Hiroshima and Nagasaki survivors. Animal studies do not seem to correlate well with effects on humans and are not very helpful. For both cancer and genetic defects, the approach to safety has been to use the linear hypothesis, which is likely to be an overestimate of the risks of low doses. Certain researchers even claim that low doses are *beneficial*. **Hormesis** is a term used to describe generally favorable biological responses to low exposures of toxins or radiation. Such low levels may help certain repair mechanisms to develop or enable cells to adapt to the effects of the low exposures. Positive effects may occur at low doses that could be a problem at high doses.

Even the linear hypothesis estimates of the risks are relatively small, and the average person is not exposed to large amounts of radiation. [link] lists average annual background radiation doses from natural and artificial sources for Australia, the United States, Germany, and world-wide averages. Cosmic rays are partially shielded by the atmosphere, and the dose depends upon altitude and latitude, but the average is about 0.40 mSv/y. A good example of the variation of cosmic radiation dose with altitude comes from the airline industry. Monitored personnel show an average of 2 mSv/y. A 12-hour flight might give you an exposure of 0.02 to 0.03 mSv.

Doses from the Earth itself are mainly due to the isotopes of uranium, thorium, and potassium, and vary greatly by location. Some places have great natural concentrations of uranium and thorium, yielding doses ten times as high as the average value. Internal doses come from foods and liquids that we ingest. Fertilizers containing phosphates have potassium and uranium. So we are all a little radioactive. Carbon-14 has about 66 Bq/kg radioactivity whereas fertilizers may have more than 3000 Bq/kg radioactivity. Medical and dental diagnostic exposures are mostly from x-rays. It should be noted that x-ray doses tend to be localized and are becoming much smaller with improved techniques. [link] shows typical doses received during various diagnostic x-ray examinations. Note the large dose from a CT scan. While CT scans only account for less than 20 percent of the x-ray procedures done today, they account for about 50 percent of the annual dose received.

Radon is usually more pronounced underground and in buildings with low air exchange with the outside world. Almost all soil contains some ^{226}Ra and ^{222}Rn, but radon is lower in mainly sedimentary soils and higher in granite soils. Thus, the exposure to the public can vary greatly, even within short distances. Radon can diffuse from the soil into homes, especially basements. The estimated exposure for ^{222}Rn is controversial. Recent studies indicate there is more radon in homes than had been realized, and it is speculated that radon may be responsible for 20 percent of lung cancers, being particularly hazardous to those who also smoke. Many countries have introduced limits on allowable radon concentrations in indoor air, often requiring the measurement of radon concentrations in a house prior to its sale. Ironically, it could be argued that the higher levels of radon exposure and their geographic variability, taken with the lack of demographic evidence of any effects, means that low-level radiation is *less* dangerous than previously thought.

Radiation Protection

Laws regulate radiation doses to which people can be exposed. The greatest occupational whole-body dose that is allowed depends upon the country and is about 20 to 50 mSv/y and is rarely reached by medical and nuclear power workers. Higher doses are allowed for the hands. Much lower doses are permitted for the reproductive organs and the fetuses of pregnant women. Inadvertent doses to the public are limited to $1/10$ of occupational doses, except for those caused by nuclear power, which cannot legally expose the public to more than $1/1000$ of the occupational limit or 0.05 mSv/y (5 mrem/y). This has been exceeded in the United States only at the time of the Three Mile Island (TMI) accident in 1979. Chernobyl is another story. Extensive monitoring with a variety of radiation detectors is performed to assure radiation safety. Increased ventilation in uranium mines has lowered the dose there to about 1 mSv/y.

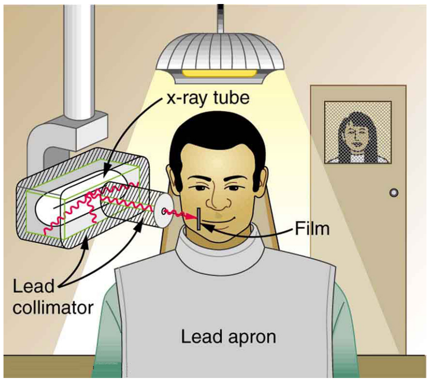
Source	**Dose (mSv/y)[footnote]** **Multiply by 100 to obtain dose in mrem/y.**			
Source	Australia	Germany	United States	World
Natural Radiation - external				

Source	**Dose (mSv/y)[footnote]** **Multiply by 100 to obtain dose in mrem/y.**			
Cosmic Rays	0.30	0.28	0.30	0.39
Soil, building materials	0.40	0.40	0.30	0.48
Radon gas	0.90	1.1	2.0	1.2
Natural Radiation - internal				
$^{40}\text{K}, {}^{14}\text{C}, {}^{226}\text{Ra}$	0.24	0.28	0.40	0.29
Medical & Dental	0.80	0.90	0.53	0.40
TOTAL	2.6	3.0	3.5	2.8

Background Radiation Sources and Average Doses

To physically limit radiation doses, we use **shielding**, increase the **distance** from a source, and limit the **time of exposure**.

[link] illustrates how these are used to protect both the patient and the dental technician when an x-ray is taken. Shielding absorbs radiation and can be provided by any material, including sufficient air. The greater the distance from the source, the more the radiation spreads out. The less time a person is exposed to a given source, the smaller is the dose received by the person. Doses from most medical diagnostics have decreased in recent years due to faster films that require less exposure time.



A lead apron is placed over the dental patient and shielding surrounds the x-ray tube to limit exposure to tissue other than the tissue that is being imaged. Fast films limit the time needed to obtain images, reducing exposure to the imaged tissue. The technician stands a few meters away behind a lead-lined door with a lead glass window, reducing her occupational exposure.

Procedure	**Effective dose (mSv)**
Chest	0.02
Dental	0.01
Skull	0.07
Leg	0.02
Mammogram	0.40
Barium enema	7.0
Upper GI	3.0
CT head	2.0
CT abdomen	10.0

Typical Doses Received During Diagnostic X-ray Exams

Problem-Solving Strategy

You need to follow certain steps for dose calculations, which are

Step 1. *Examine the situation to determine that a person is exposed to ionizing radiation.*

Step 2. *Identify exactly what needs to be determined in the problem (identify the unknowns).* The most straightforward problems ask for a dose calculation.

Step 3. *Make a list of what is given or can be inferred from the problem as stated (identify the knowns).* Look for information on the type of radiation, the energy per event, the activity, and the mass of tissue affected.

Step 4. *For dose calculations, you need to determine the energy deposited.* This may take one or more steps, depending on the given information.

Step 5. *Divide the deposited energy by the mass of the affected tissue.* Use units of joules for energy and kilograms for mass. If a dose in Sv is involved, use the definition that $1\text{ Sv}=1\text{ J/kg}$.

Step 6. *If a dose in mSv is involved, determine the RBE (QF) of the radiation.* Recall that $1\text{ mSv}=1\text{ mGy}\times\text{RBE}\text{ (or }1\text{ rem}=1\text{ rad}\times\text{RBE)}$.

Step 7. *Check the answer to see if it is reasonable: Does it make sense?* The dose should be consistent with the numbers given in the text for diagnostic, occupational, and therapeutic exposures.

Example:
Dose from Inhaled Plutonium

Calculate the dose in rem/y for the lungs of a weapons plant employee who inhales and retains an activity of 1.00 μCi of ^{239}Pu in an accident. The mass of affected lung tissue is 2.00 kg, the plutonium decays by emission of a 5.23-MeV α particle, and you may assume the higher value of the RBE for α s from [link].

Strategy

Dose in rem is defined by $1\text{ rad}=0.01\text{ J/kg}$ and $\text{rem}=\text{rad}\times\text{RBE}$. The energy deposited is divided by the mass of tissue affected and then multiplied by the RBE. The latter two quantities are given, and so the main task in this example will be to find the energy deposited in one year. Since the activity of the source is given, we can calculate the number of decays, multiply by the energy per decay, and convert MeV to joules to get the total energy.

Solution

The activity $R = 1.00\ \text{µCi} = 3.70 \times 10^4\ \text{Bq} = 3.70 \times 10^4$ decays/s. So, the number of decays per year is obtained by multiplying by the number of seconds in a year:

Equation:

$$\left(3.70 \times 10^4\ \text{decays/s}\right)\left(3.16 \times 10^7\ \text{s}\right) = 1.17 \times 10^{12}\ \text{decays.}$$

Thus, the ionizing energy deposited per year is

Equation:

$$E = \left(1.17 \times 10^{12}\ \text{decays}\right)\left(5.23\ \text{MeV/decay}\right) \times \left(\frac{1.60 \times 10^{-13}\ \text{J}}{\text{MeV}}\right) = 0.978\ \text{J.}$$

Dividing by the mass of the affected tissue gives

Equation:

$$\frac{E}{\text{mass}} = \frac{0.978\ \text{J}}{2.00\ \text{kg}} = 0.489\ \text{J/kg.}$$

One Gray is 1.00 J/kg, and so the dose in Gy is

Equation:

$$\text{dose in Gy} = \frac{0.489\ \text{J/kg}}{1.00\ \text{(J/kg)/Gy}} = 0.489\ \text{Gy.}$$

Now, the dose in Sv is

Equation:

$$\text{dose in Sv} = \text{Gy} \times \text{RBE}$$

Equation:

$$= (0.489\ \text{Gy})(20) = 9.8\ \text{Sv.}$$

Discussion

First note that the dose is given to two digits, because the RBE is (at best) known only to two digits. By any standard, this yearly radiation dose is high and will have a devastating effect on the health of the worker. Worse yet, plutonium has a long radioactive half-life and is not readily eliminated by the body, and so it will remain in the lungs. Being an α emitter makes the effects 10 to 20 times worse than the same ionization produced by β s, γ rays, or x-rays. An activity of 1.00 μCi is created by only 16 μg of ^{239}Pu (left as an end-of-chapter problem to verify), partly justifying claims that plutonium is the most toxic substance known. Its actual hazard depends on how likely it is to be spread out among a large population and then ingested. The Chernobyl disaster's deadly legacy, for example, has nothing to do with the plutonium it put into the environment.

Risk versus Benefit

Medical doses of radiation are also limited. Diagnostic doses are generally low and have further lowered with improved techniques and faster films. With the possible exception of routine dental x-rays, radiation is used diagnostically only when needed so that the low risk is justified by the benefit of the diagnosis. Chest x-rays give the lowest doses—about 0.1 mSv to the tissue affected, with less than 5 percent scattering into tissues that are not directly imaged. Other x-ray procedures range upward to about 10 mSv in a CT scan, and about 5 mSv (0.5 rem) per dental x-ray, again both only affecting the tissue imaged. Medical images with radiopharmaceuticals give doses ranging from 1 to 5 mSv, usually localized. One exception is the thyroid scan using ^{131}I. Because of its relatively long half-life, it exposes the thyroid to about 0.75 Sv. The isotope ^{123}I is more difficult to produce, but its short half-life limits thyroid exposure to about 15 mSv.

Note:
PhET Explorations: Alpha Decay
Watch alpha particles escape from a polonium nucleus, causing radioactive alpha decay. See how random decay times relate to the half life.

Alpha Decay.

Section Summary

- The biological effects of ionizing radiation are due to two effects it has on cells: interference with cell reproduction, and destruction of cell function.
- A radiation dose unit called the rad is defined in terms of the ionizing energy deposited per kilogram of tissue:
  Equation:
  $$1\text{ rad}=0.01\text{ J/kg}.$$
- The SI unit for radiation dose is the gray (Gy), which is defined to be $1\text{ Gy}=1\text{ J/kg}=100\text{ rad}.$
- To account for the effect of the type of particle creating the ionization, we use the relative biological effectiveness (RBE) or quality factor (QF) given in [link] and define a unit called the roentgen equivalent man (rem) as

Equation:

$$\text{rem} = \text{rad} \times \text{RBE}.$$

- Particles that have short ranges or create large ionization densities have RBEs greater than unity. The SI equivalent of the rem is the sievert (Sv), defined to be
  Equation:

  $$\text{Sv} = \text{Gy} \times \text{RBE and } 1\ \text{Sv} = 100\ \text{rem}.$$

- Whole-body, single-exposure doses of 0.1 Sv or less are low doses while those of 0.1 to 1 Sv are moderate, and those over 1 Sv are high doses. Some immediate radiation effects are given in [link]. Effects due to low doses are not observed, but their risk is assumed to be directly proportional to those of high doses, an assumption known as the linear hypothesis. Long-term effects are cancer deaths at the rate of $10/10^6\ \text{rem}\cdot\text{y}$and genetic defects at roughly one-third this rate. Background radiation doses and sources are given in [link]. World-wide average radiation exposure from natural sources, including radon, is about 3 mSv, or 300 mrem. Radiation protection utilizes shielding, distance, and time to limit exposure.

Conceptual Questions

Exercise:

Problem:

Isotopes that emit α radiation are relatively safe outside the body and exceptionally hazardous inside. Yet those that emit γ radiation are hazardous outside and inside. Explain why.

Exercise:

Problem:

Why is radon more closely associated with inducing lung cancer than other types of cancer?

Exercise:

Problem:

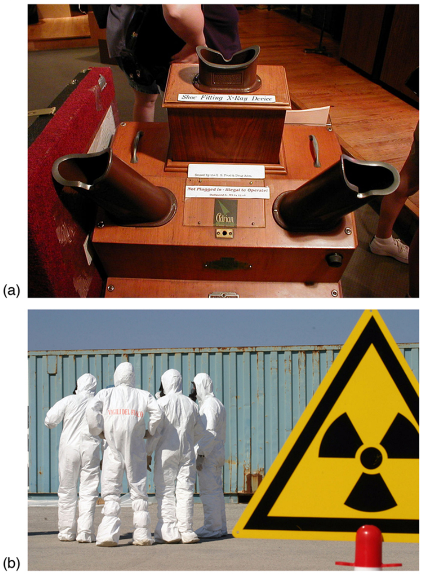
The RBE for low-energy βs is 1.7, whereas that for higher-energy βs is only 1. Explain why, considering how the range of radiation depends on its energy.

Exercise:

Problem:

Which methods of radiation protection were used in the device shown in the first photo in [link]? Which were used in the situation shown in the second photo?

(a)



(a) This x-ray fluorescence machine is one of the thousands used in shoe stores to produce images of feet as a check on the fit of shoes. They are unshielded and remain on as long as the feet are in them, producing doses much greater than medical images. Children were fascinated with them. These machines were used in shoe stores until laws preventing such unwarranted radiation exposure were enacted in the 1950s. (credit: Andrew Kuchling) (b) Now that we know the effects of exposure to

radioactive material,
safety is a priority.
(credit: U.S. Navy)

Exercise:

Problem:

What radioisotope could be a problem in homes built of cinder blocks made from uranium mine tailings? (This is true of homes and schools in certain regions near uranium mines.)

Exercise:

Problem:

Are some types of cancer more sensitive to radiation than others? If so, what makes them more sensitive?

Exercise:

Problem:

Suppose a person swallows some radioactive material by accident. What information is needed to be able to assess possible damage?

Problems & Exercises

Exercise:

Problem:

What is the dose in mSv for: (a) a 0.1 Gy x-ray? (b) 2.5 mGy of neutron exposure to the eye? (c) 1.5 mGy of α exposure?

Solution:

(a) 100 mSv

(b) 80 mSv

(c) ~30 mSv

Exercise:

Problem:

Find the radiation dose in Gy for: (a) A 10-mSv fluoroscopic x-ray series. (b) 50 mSv of skin exposure by an α emitter. (c) 160 mSv of β^- and γ rays from the ^{40}K in your body.

Exercise:

Problem:

How many Gy of exposure is needed to give a cancerous tumor a dose of 40 Sv if it is exposed to α activity?

Solution:

~2 Gy

Exercise:

Problem:

What is the dose in Sv in a cancer treatment that exposes the patient to 200 Gy of γ rays?

Exercise:

Problem:

One half the γ rays from ^{99m}Tc are absorbed by a 0.170-mm-thick lead shielding. Half of the γ rays that pass through the first layer of lead are absorbed in a second layer of equal thickness. What thickness of lead will absorb all but one in 1000 of these γ rays?

Solution:

1.69 mm

Exercise:

Problem:

A plumber at a nuclear power plant receives a whole-body dose of 30 mSv in 15 minutes while repairing a crucial valve. Find the radiation-induced yearly risk of death from cancer and the chance of genetic defect from this maximum allowable exposure.

Exercise:

Problem:

In the 1980s, the term picowave was used to describe food irradiation in order to overcome public resistance by playing on the well-known safety of microwave radiation. Find the energy in MeV of a photon having a wavelength of a picometer.

Solution:

1.24 MeV

Exercise:

Problem: Find the mass of ^{239}Pu that has an activity of 1.00 μCi.

Glossary

gray (Gy)
: the SI unit for radiation dose which is defined to be $1\text{ Gy} = 1\text{ J/kg} = 100\text{ rad}$

linear hypothesis
: assumption that risk is directly proportional to risk from high doses

rad
: the ionizing energy deposited per kilogram of tissue

sievert
: the SI equivalent of the rem

relative biological effectiveness (RBE)
: a number that expresses the relative amount of damage that a fixed amount of ionizing radiation of a given type can inflict on biological tissues

quality factor
: same as relative biological effectiveness

roentgen equivalent man (rem)
: a dose unit more closely related to effects in biological tissue

low dose
: a dose less than 100 mSv (10 rem)

moderate dose
: a dose from 0.1 Sv to 1 Sv (10 to 100 rem)

high dose

a dose greater than 1 Sv (100 rem)

hormesis

a term used to describe generally favorable biological responses to low exposures of toxins or radiation

shielding

a technique to limit radiation exposure

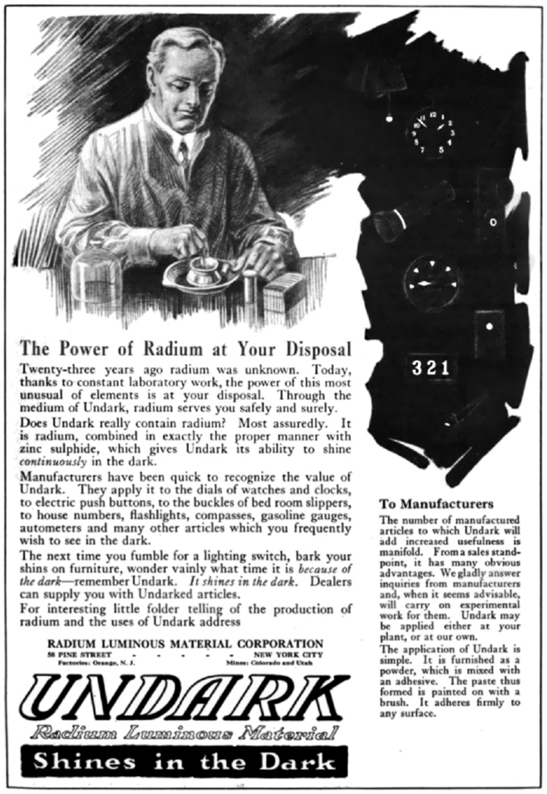
Therapeutic Uses of Ionizing Radiation

- Explain the concept of radiotherapy and list typical doses for cancer therapy.

Therapeutic applications of ionizing radiation, called radiation therapy or **radiotherapy**, have existed since the discovery of x-rays and nuclear radioactivity. Today, radiotherapy is used almost exclusively for cancer therapy, where it saves thousands of lives and improves the quality of life and longevity of many it cannot save. Radiotherapy may be used alone or in combination with surgery and chemotherapy (drug treatment) depending on the type of cancer and the response of the patient. A careful examination of all available data has established that radiotherapy's beneficial effects far outweigh its long-term risks.

Medical Application

The earliest uses of ionizing radiation on humans were mostly harmful, with many at the level of snake oil as seen in [link]. Radium-doped cosmetics that glowed in the dark were used around the time of World War I. As recently as the 1950s, radon mine tours were promoted as healthful and rejuvenating—those who toured were exposed but gained no benefits. Radium salts were sold as health elixirs for many years. The gruesome death of a wealthy industrialist, who became psychologically addicted to the brew, alerted the unsuspecting to the dangers of radium salt elixirs. Most abuses finally ended after the legislation in the 1950s.



The Power of Radium at Your Disposal

Twenty-three years ago radium was unknown. Today, thanks to constant laboratory work, the power of this most unusual of elements is at your disposal. Through the medium of Undark, radium serves you safely and surely.

Does Undark really contain radium? Most assuredly. It is radium, combined in exactly the proper manner with zinc sulphide, which gives Undark its ability to shine *continuously* in the dark.

Manufacturers have been quick to recognize the value of Undark. They apply it to the dials of watches and clocks, to electric push buttons, to the buckles of bed room slippers, to house numbers, flashlights, compasses, gasoline gauges, autometers and many other articles which you frequently wish to see in the dark.

The next time you fumble for a lighting switch, bark your shins on furniture, wonder vainly what time it is *because of the dark*—remember Undark. *It shines in the dark.* Dealers can supply you with Undarked articles.

For interesting little folder telling of the production of radium and the uses of Undark address

RADIUM LUMINOUS MATERIAL CORPORATION
58 PINE STREET - - - - - NEW YORK CITY
Factories: Orange, N. J. Mines: Colorado and Utah

UNDARK
Radium Luminous Material
Shines in the Dark

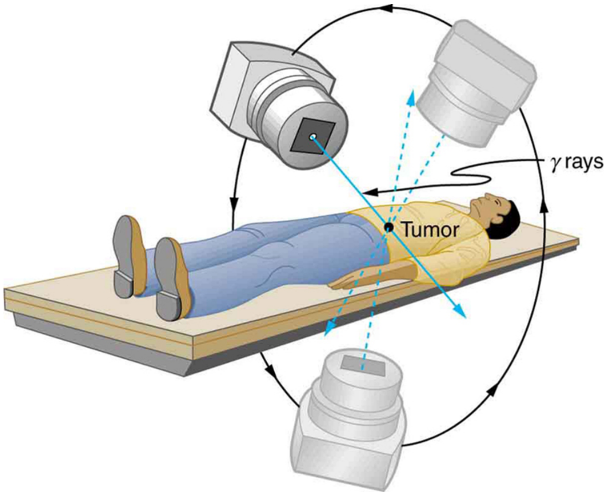
To Manufacturers

The number of manufactured articles to which Undark will add increased usefulness is manifold. From a sales standpoint, it has many obvious advantages. We gladly answer inquiries from manufacturers and, when it seems advisable, will carry on experimental work for them. Undark may be applied either at your plant, or at our own.

The application of Undark is simple. It is furnished as a powder, which is mixed with an adhesive. The paste thus formed is painted on with a brush. It adheres firmly to any surface.

The properties of radiation were once touted for far more than its modern use in cancer therapy. Until 1932, radium was advertised for a variety of uses, often with tragic results. (credit: Struthious Bandersnatch.)

Radiotherapy is effective against cancer because cancer cells reproduce rapidly and, consequently, are more sensitive to radiation. The central problem in radiotherapy is to make the dose for cancer cells as high as possible while limiting the dose for normal cells. The ratio of abnormal cells killed to normal cells killed is called the **therapeutic ratio**, and all radiotherapy techniques are designed to enhance this ratio. Radiation can be concentrated in cancerous tissue by a number of techniques. One of the most prevalent techniques for well-defined tumors is a geometric technique shown in [link]. A narrow beam of radiation is passed through the patient from a variety of directions with a common crossing point in the tumor. This concentrates the dose in the tumor while spreading it out over a large volume of normal tissue. The external radiation can be x-rays, ^{60}Co γ rays, or ionizing-particle beams produced by accelerators. Accelerator-produced beams of neutrons, π-mesons, and heavy ions such as nitrogen nuclei have been employed, and these can be quite effective. These particles have larger QFs or RBEs and sometimes can be better localized, producing a greater therapeutic ratio. But accelerator radiotherapy is much more expensive and less frequently employed than other forms.



The ^{60}Co source of γ-radiation is rotated around the patient so that the common crossing point is in the tumor, concentrating the dose there. This geometric technique works for well-defined tumors.

Another form of radiotherapy uses chemically inert radioactive implants. One use is for prostate cancer. Radioactive seeds (about 40 to 100 and the size of a grain of rice) are placed in the prostate region. The isotopes used are usually ^{135}I (6-month half life) or ^{103}Pd (3-month half life). Alpha emitters have the dual advantages of a large QF and a small range for better localization.

Radiopharmaceuticals are used for cancer therapy when they can be localized well enough to produce a favorable therapeutic ratio. Thyroid cancer is commonly treated utilizing radioactive iodine. Thyroid cells concentrate iodine, and cancerous thyroid cells are more aggressive in doing this. An ingenious use of radiopharmaceuticals in cancer therapy tags antibodies with radioisotopes. Antibodies produced by a patient to combat his cancer are extracted, cultured, loaded with a radioisotope, and then returned to the patient. The antibodies are concentrated almost entirely in the tissue they developed to fight, thus localizing the radiation in abnormal tissue. The therapeutic ratio can be quite high for short-range radiation. There is, however, a significant dose for organs that eliminate radiopharmaceuticals from the body, such as the liver, kidneys, and bladder. As with most radiotherapy, the technique is limited by the tolerable amount of damage to the normal tissue.

[link] lists typical therapeutic doses of radiation used against certain cancers. The doses are large, but not fatal because they are localized and spread out in time. Protocols for treatment vary with the type of cancer and the condition and response of the patient. Three to five 200-rem treatments per week for a period of several weeks is typical. Time between treatments allows the body to repair normal tissue. This effect occurs because damage is concentrated in the abnormal tissue, and the abnormal tissue is more sensitive to radiation. Damage to normal tissue limits the doses. You will note that the greatest doses are given to any tissue that is not rapidly reproducing, such as in the adult brain. Lung cancer, on the other end of the scale, cannot ordinarily be cured with radiation because of the sensitivity of lung tissue and blood to radiation. But radiotherapy for lung cancer does alleviate symptoms and prolong life and is therefore justified in some cases.

Type of Cancer	**Typical dose (Sv)**
Lung	10–20
Hodgkin’s disease	40–45
Skin	40–50
Ovarian	50–75
Breast	50–80+
Brain	80+
Neck	80+
Bone	80+
Soft tissue	80+
Thyroid	80+

Cancer Radiotherapy

Finally, it is interesting to note that chemotherapy employs drugs that interfere with cell division and is, thus, also effective against cancer. It also has almost the same side effects, such as nausea and hair loss, and risks, such as the inducement of another cancer.

Section Summary

- Radiotherapy is the use of ionizing radiation to treat ailments, now limited to cancer therapy.
- The sensitivity of cancer cells to radiation enhances the ratio of cancer cells killed to normal cells killed, which is called the therapeutic ratio.

- Doses for various organs are limited by the tolerance of normal tissue for radiation. Treatment is localized in one region of the body and spread out in time.

Conceptual Questions

Exercise:

Problem:

Radiotherapy is more likely to be used to treat cancer in elderly patients than in young ones. Explain why. Why is radiotherapy used to treat young people at all?

Problems & Exercises

Exercise:

Problem:

A beam of 168-MeV nitrogen nuclei is used for cancer therapy. If this beam is directed onto a 0.200-kg tumor and gives it a 2.00-Sv dose, how many nitrogen nuclei were stopped? (Use an RBE of 20 for heavy ions.)

Solution:

7.44×10^{8}

Exercise:

Problem:

(a) If the average molecular mass of compounds in food is 50.0 g, how many molecules are there in 1.00 kg of food? (b) How many ion pairs are created in 1.00 kg of food, if it is exposed to 1000 Sv and it takes 32.0 eV to create an ion pair? (c) Find the ratio of ion pairs to molecules. (d) If these ion pairs recombine into a distribution of 2000 new compounds, how many parts per billion is each?

Exercise:

Problem:

Calculate the dose in Sv to the chest of a patient given an x-ray under the following conditions. The x-ray beam intensity is 1.50 W/m^2, the area of the chest exposed is 0.0750 m^2, 35.0% of the x-rays are absorbed in 20.0 kg of tissue, and the exposure time is 0.250 s.

Solution:

$4.92 \times 10^{-4}\text{ Sv}$

Exercise:

Problem:

(a) A cancer patient is exposed to γ rays from a 5000-Ci ^{60}Co transillumination unit for 32.0 s. The γ rays are collimated in such a manner that only 1.00% of them strike the patient. Of those, 20.0% are absorbed in a tumor having a mass of 1.50 kg. What is the dose in rem to the tumor, if the average γ energy per decay is 1.25 MeV? None of the β s from the decay reach the patient. (b) Is the dose consistent with stated therapeutic doses?

Exercise:

Problem:

What is the mass of ^{60}Co in a cancer therapy transillumination unit containing 5.00 kCi of ^{60}Co?

Solution:

4.43 g

Exercise:

Problem:

Large amounts of ^{65}Zn are produced in copper exposed to accelerator beams. While machining contaminated copper, a physicist ingests $50.0\ \mu\text{Ci}$ of ^{65}Zn. Each ^{65}Zn decay emits an average γ-ray energy of 0.550 MeV, 40.0% of which is absorbed in the scientist’s 75.0-kg body. What dose in mSv is caused by this in one day?

Exercise:

Problem:

Naturally occurring ^{40}K is listed as responsible for 16 mrem/y of background radiation. Calculate the mass of ^{40}K that must be inside the 55-kg body of a woman to produce this dose. Each ^{40}K decay emits a 1.32-MeV β, and 50% of the energy is absorbed inside the body.

Solution:

0.010 g

Exercise:

Problem:

(a) Background radiation due to ^{226}Ra averages only 0.01 mSv/y, but it can range upward depending on where a person lives. Find the mass of ^{226}Ra in the 80.0-kg body of a man who receives a dose of 2.50-mSv/y from it, noting that each ^{226}Ra decay emits a 4.80-MeV α particle. You may neglect dose due to daughters and assume a constant amount, evenly distributed due to balanced ingestion and bodily elimination. (b) Is it surprising that such a small mass could cause a measurable radiation dose? Explain.

Exercise:

Problem:

The annual radiation dose from ^{14}C in our bodies is 0.01 mSv/y. Each ^{14}C decay emits a β^{-} averaging 0.0750 MeV. Taking the fraction of ^{14}C to be 1.3×10^{-12} N of normal ^{12}C, and assuming the body is 13% carbon, estimate the fraction of the decay energy absorbed. (The rest escapes, exposing those close to you.)

Solution:

95%

Exercise:

Problem:

If everyone in Australia received an extra 0.05 mSv per year of radiation, what would be the increase in the number of cancer deaths per year? (Assume that time had elapsed for the effects to become apparent.) Assume that there are 200×10^{-4} deaths per Sv of radiation per year. What percent of the actual number of cancer deaths recorded is this?

Glossary

radiotherapy
    the use of ionizing radiation to treat ailments

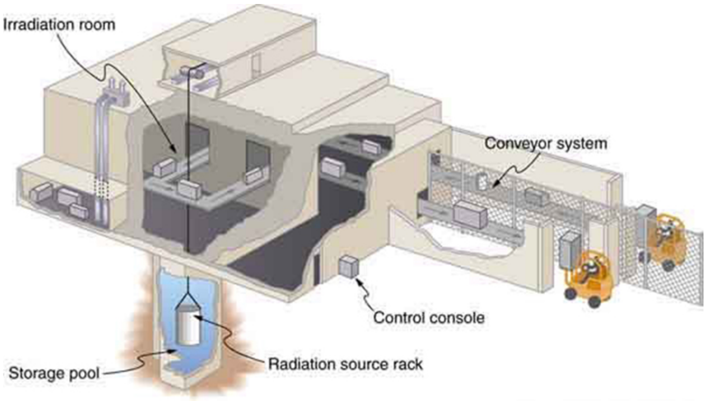
therapeutic ratio
    the ratio of abnormal cells killed to normal cells killed

Food Irradiation

- Define food irradiation low dose, and free radicals.

Ionizing radiation is widely used to sterilize medical supplies, such as bandages, and consumer products, such as tampons. Worldwide, it is also used to irradiate food, an application that promises to grow in the future. **Food irradiation** is the treatment of food with ionizing radiation. It is used to reduce pest infestation and to delay spoilage and prevent illness caused by microorganisms. Food irradiation is controversial. Proponents see it as superior to pasteurization, preservatives, and insecticides, supplanting dangerous chemicals with a more effective process. Opponents see its safety as unproven, perhaps leaving worse toxic residues as well as presenting an environmental hazard at treatment sites. In developing countries, food irradiation might increase crop production by 25.0% or more, and reduce food spoilage by a similar amount. It is used chiefly to treat spices and some fruits, and in some countries, red meat, poultry, and vegetables. Over 40 countries have approved food irradiation at some level.

Food irradiation exposes food to large doses of γ rays, x-rays, or electrons. These photons and electrons induce no nuclear reactions and thus create *no residual radioactivity*. (Some forms of ionizing radiation, such as neutron irradiation, cause residual radioactivity. These are not used for food irradiation.) The γ source is usually ^{60}Co or ^{137}Cs, the latter isotope being a major by-product of nuclear power. Cobalt-60 γ rays average 1.25 MeV, while those of ^{137}Cs are 0.67 MeV and are less penetrating. X-rays used for food irradiation are created with voltages of up to 5 million volts and, thus, have photon energies up to 5 MeV. Electrons used for food irradiation are accelerated to energies up to 10 MeV. The higher the energy per particle, the more penetrating the radiation is and the more ionization it can create. [link] shows a typical γ-irradiation plant.



A food irradiation plant has a conveyor system to pass items through an intense radiation field behind thick shielding walls. The γ source is lowered into a deep pool of water for safe storage when not in use. Exposure times of up to an hour expose food to doses up to 10^4 Gy.

Owing to the fact that food irradiation seeks to destroy organisms such as insects and bacteria, much larger doses than those fatal to humans must be applied. Generally, the simpler the organism, the more radiation it can tolerate. (Cancer cells are a partial exception, because they are rapidly reproducing and, thus, more sensitive.) Current licensing allows up to 1000 Gy to be applied to fresh fruits and vegetables, called a *low dose* in food irradiation. Such a dose is enough to prevent or reduce the growth of many microorganisms, but about 10,000 Gy is needed to kill salmonella, and even more is needed to kill fungi. Doses greater than 10,000 Gy are considered to be high doses in food irradiation and product sterilization.

The effectiveness of food irradiation varies with the type of food. Spices and many fruits and vegetables have dramatically longer shelf lives. These also show no degradation in taste and no loss of food value or vitamins. If not for the mandatory labeling, such foods subjected to low-level irradiation (up to 1000 Gy) could not be distinguished from untreated foods in quality.

However, some foods actually spoil faster after irradiation, particularly those with high water content like lettuce and peaches. Others, such as milk, are given a noticeably unpleasant taste. High-level irradiation produces significant and chemically measurable changes in foods. It produces about a 15% loss of nutrients and a 25% loss of vitamins, as well as some change in taste. Such losses are similar to those that occur in ordinary freezing and cooking.

How does food irradiation work? Ionization produces a random assortment of broken molecules and ions, some with unstable oxygen- or hydrogen-containing molecules known as **free radicals**. These undergo rapid chemical reactions, producing perhaps four or five thousand different compounds called **radiolytic products**, some of which make cell function impossible by breaking cell membranes, fracturing DNA, and so on. How safe is the food afterward? Critics argue that the radiolytic products present a lasting hazard, perhaps being carcinogenic. However, the safety of irradiated food is not known precisely. We do know that low-level food irradiation produces no compounds in amounts that can be measured chemically. This is not surprising, since trace amounts of several thousand compounds may be created. We also know that there have been no observable negative short-term effects on consumers. Long-term effects may show up if large number of people consume large quantities of irradiated food, but no effects have appeared due to the small amounts of irradiated food that are consumed regularly. The case for safety is supported by testing of animal diets that were irradiated; no transmitted genetic effects have been observed. Food irradiation (at least up to a million rad) has been endorsed by the World Health Organization and the UN Food and Agricultural Organization. Finally, the hazard to consumers, if it exists, must be weighed against the benefits in food production and preservation. It must also be weighed against the very real hazards of existing insecticides and food preservatives.

Section Summary

- Food irradiation is the treatment of food with ionizing radiation.
- Irradiating food can destroy insects and bacteria by creating free radicals and radiolytic products that can break apart cell membranes.

- Food irradiation has produced no observable negative short-term effects for humans, but its long-term effects are unknown.

Conceptual Questions

Exercise:

Problem:

Does food irradiation leave the food radioactive? To what extent is the food altered chemically for low and high doses in food irradiation?

Exercise:

Problem:

Compare a low dose of radiation to a human with a low dose of radiation used in food treatment.

Exercise:

Problem:

Suppose one food irradiation plant uses a ^{137}Cs source while another uses an equal activity of ^{60}Co. Assuming equal fractions of the γ rays from the sources are absorbed, why is more time needed to get the same dose using the ^{137}Cs source?

Glossary

food irradiation
: treatment of food with ionizing radiation

free radicals
: ions with unstable oxygen- or hydrogen-containing molecules

radiolytic products
: compounds produced due to chemical reactions of free radicals

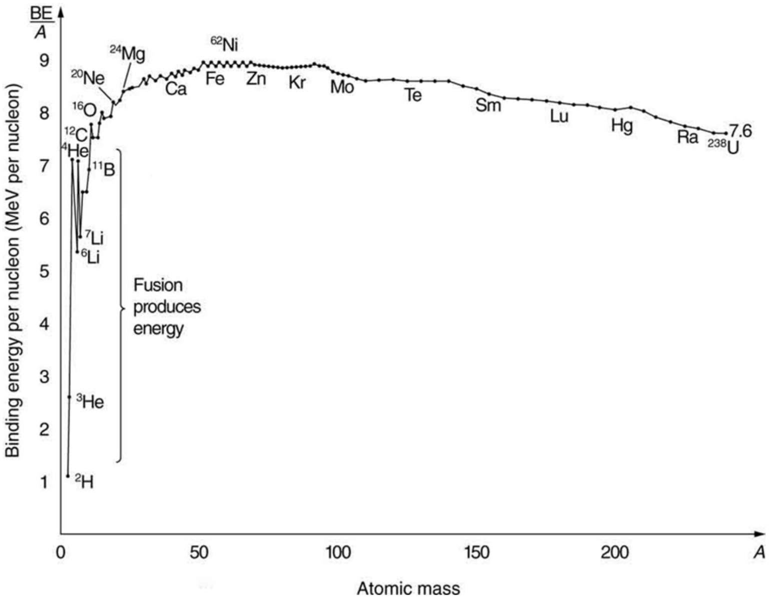
Fusion

- Define nuclear fusion.
- Discuss processes to achieve practical fusion energy generation.

While basking in the warmth of the summer sun, a student reads of the latest breakthrough in achieving sustained thermonuclear power and vaguely recalls hearing about the cold fusion controversy. The three are connected. The Sun’s energy is produced by nuclear fusion (see [link]). Thermonuclear power is the name given to the use of controlled nuclear fusion as an energy source. While research in the area of thermonuclear power is progressing, high temperatures and containment difficulties remain. The cold fusion controversy centered around unsubstantiated claims of practical fusion power at room temperatures.

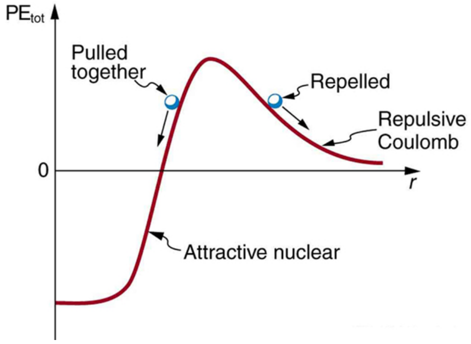
The Sun’s energy is produced by nuclear fusion. (credit: Spiralz)

Nuclear fusion is a reaction in which two nuclei are combined, or *fused*, to form a larger nucleus. We know that all nuclei have less mass than the sum of the masses of the protons and neutrons that form them. The missing mass times c^2 equals the binding energy of the nucleus—the greater the binding energy, the greater the missing mass. We also know that BE/A, the binding energy per nucleon, is greater for medium-mass nuclei and has a maximum at Fe (iron). This means that if two low-mass nuclei can be fused together to form a larger nucleus, energy can be released. The larger nucleus has a greater binding energy and less mass per nucleon than the two that combined. Thus mass is destroyed in the fusion reaction, and energy is released (see [link]). On average, fusion of low-mass nuclei releases energy, but the details depend on the actual nuclides involved.



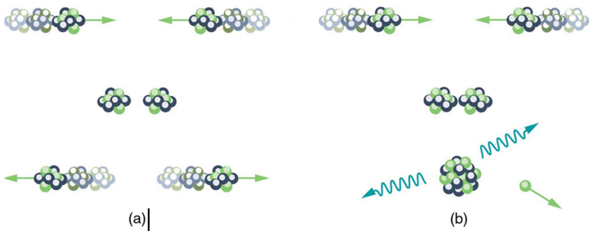
Fusion of light nuclei to form medium-mass nuclei destroys mass, because BE/A is greater for the product nuclei. The larger BE/A is, the less mass per nucleon, and so mass is converted to energy and released in these fusion reactions.

The major obstruction to fusion is the Coulomb repulsion between nuclei. Since the attractive nuclear force that can fuse nuclei together is short ranged, the repulsion of like positive charges must be overcome to get nuclei close enough to induce fusion. [link] shows an approximate graph of the potential energy between two nuclei as a function of the distance between their centers. The graph is analogous to a hill with a well in its center. A ball rolled from the right must have enough kinetic energy to get over the hump before it falls into the deeper well with a net gain in energy. So it is with fusion. If the nuclei are given enough kinetic energy to overcome the electric potential energy due to repulsion, then they can combine, release energy, and fall into a deep well. One way to accomplish this is to heat fusion fuel to high temperatures so that the kinetic energy of thermal motion is sufficient to get the nuclei together.



Potential energy between two light nuclei graphed as a function of distance between them. If the nuclei have enough kinetic energy to get over the Coulomb repulsion hump, they combine, release energy, and drop into a deep attractive well. Tunneling through the barrier is important in practice. The greater the kinetic energy and the higher the particles get up the barrier (or the lower the barrier), the more likely the tunneling.

You might think that, in the core of our Sun, nuclei are coming into contact and fusing. However, in fact, temperatures on the order of 10^8K are needed to actually get the nuclei in contact, exceeding the core temperature of the Sun. Quantum mechanical tunneling is what makes fusion in the Sun possible, and tunneling is an important process in most other practical applications of fusion, too. Since the probability of tunneling is extremely sensitive to barrier height and width, increasing the temperature greatly increases the rate of fusion. The closer reactants get to one another, the more likely they are to fuse (see [link]). Thus most fusion in the Sun and other stars takes place at their centers, where temperatures are highest. Moreover, high temperature is needed for thermonuclear power to be a practical source of energy.



(a) Two nuclei heading toward each other slow down, then stop, and then fly away without touching or fusing. (b) At higher energies, the two nuclei approach close enough for fusion via tunneling. The probability of tunneling increases as they approach, but they do not have to touch for the reaction to occur.

The Sun produces energy by fusing protons or hydrogen nuclei ${}^{1}\text{H}$ (by far the Sun's most abundant nuclide) into helium nuclei ${}^{4}\text{He}$. The principal sequence of fusion reactions forms what is called the **proton-proton cycle**:

Equation:

$${}^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}} \qquad (0.42\text{ MeV})$$

Equation:

$${}^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma \qquad (5.49\text{ MeV})$$

Equation:

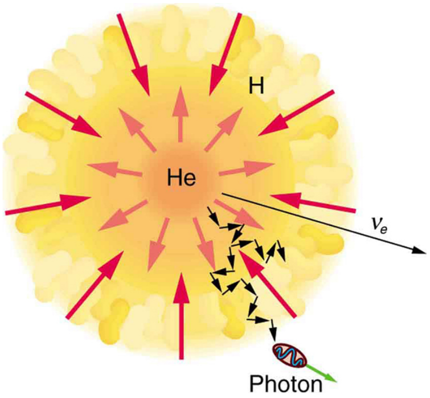
$${}^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H} \qquad (12.86\text{ MeV})$$

where e^{+} stands for a positron and v_{e} is an electron neutrino. (The energy in parentheses is *released* by the reaction.) Note that the first two reactions must occur twice for the third to be possible, so that the cycle consumes six protons (${}^{1}\text{H}$) but gives back two. Furthermore, the two positrons produced will find two electrons and annihilate to form four more γ rays, for a total of six. The overall effect of the cycle is thus

Equation:

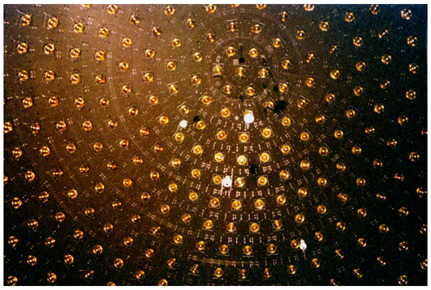
$$2e^{-} + 4{}^{1}\text{H} \rightarrow {}^{4}\text{He} + 2\text{v}_{\text{e}} + 6\gamma \qquad (26.7\text{ MeV})$$

where the 26.7 MeV includes the annihilation energy of the positrons and electrons and is distributed among all the reaction products. The solar interior is dense, and the reactions occur deep in the Sun where temperatures are highest. It takes about 32,000 years for the energy to diffuse to the surface and radiate away. However, the neutrinos escape the Sun in less than two seconds, carrying their energy with them, because they interact so weakly that the Sun is transparent to them. Negative feedback in the Sun acts as a thermostat to regulate the overall energy output. For instance, if the interior of the Sun becomes hotter than normal, the reaction rate increases, producing energy that expands the interior. This cools it and lowers the reaction rate. Conversely, if the interior becomes too cool, it contracts, increasing the temperature and reaction rate (see [link]). Stars like the Sun are stable for billions of years, until a significant fraction of their hydrogen has been depleted. What happens then is discussed in Introduction to Frontiers of Physics .

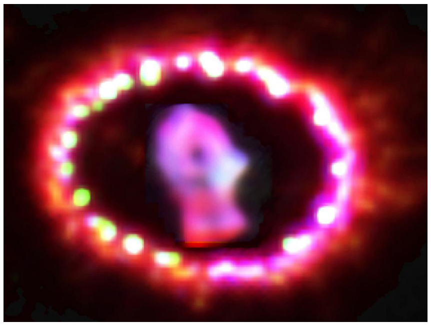


Nuclear fusion in the Sun converts hydrogen nuclei into helium; fusion occurs primarily at the boundary of the helium core, where temperature is highest and sufficient hydrogen remains. Energy released diffuses slowly to the surface, with the exception of neutrinos, which escape immediately. Energy production remains stable because of negative feedback effects.

Theories of the proton-proton cycle (and other energy-producing cycles in stars) were pioneered by the German-born, American physicist Hans Bethe (1906–2005), starting in 1938. He was awarded the 1967 Nobel Prize in physics for this work, and he has made many other contributions to physics and society. Neutrinos produced in these cycles escape so readily that they provide us an excellent means to test these theories and study stellar interiors. Detectors have been constructed and operated for more than four decades now to measure solar neutrinos (see [link]). Although solar neutrinos are detected and neutrinos were observed from Supernova 1987A ([link]), too few solar neutrinos were observed to be consistent with predictions of solar energy production. After many years, this solar neutrino problem was resolved with a blend of theory and experiment that showed that the neutrino does indeed have mass. It was also found that there are three types of neutrinos, each associated with a different type of nuclear decay.



This array of photomultiplier tubes is part of the large solar neutrino detector at the Fermi National Accelerator Laboratory in Illinois. In these experiments, the neutrinos interact with heavy water and produce flashes of light, which are detected by the photomultiplier tubes. In spite of its size and the huge flux of neutrinos that strike it, very few are detected each day since they interact so weakly. This, of course, is the same reason they escape the Sun so readily. (credit: Fred Ullrich)



Supernovas are the source of elements heavier than iron. Energy released powers nucleosynthesis. Spectroscopic analysis of the ring of material ejected by Supernova 1987A observable in the southern hemisphere, shows evidence of heavy elements. The study of this supernova also provided indications that neutrinos might have mass. (credit: NASA, ESA, and P. Challis)

The proton-proton cycle is not a practical source of energy on Earth, in spite of the great abundance of hydrogen (^1H). The reaction $^1\text{H}+{}^1\text{H}\to{}^2\text{H}+e^++v_\text{e}$ has a very low probability of occurring. (This is why our Sun will last for about ten billion years.) However, a number of other fusion reactions are easier to induce. Among them are:

Equation:

$$^2\text{H}+{}^2\text{H}\to{}^3\text{H}+{}^1\text{H}\qquad(4.03\text{ MeV})$$

Equation:

$$^2\text{H}+{}^2\text{H}\to{}^3\text{He}+n\qquad(3.27\text{ MeV})$$

Equation:

$$^2\text{H}+{}^3\text{H}\to{}^4\text{He}+n\qquad(17.59\text{ MeV})$$

Equation:

$$^2\text{H}+{}^2\text{H}\to{}^4\text{He}+\gamma\qquad(23.85\text{ MeV}).$$

Deuterium (^2H) is about 0.015% of natural hydrogen, so there is an immense amount of it in sea water alone. In addition to an abundance of deuterium fuel, these fusion reactions produce large energies per reaction (in parentheses), but they do not produce much radioactive waste. Tritium (^3H) is radioactive, but it is consumed as a fuel (the reaction $^2\text{H}+{}^3\text{H}\to{}^4\text{He}+n$), and the neutrons and γs can be shielded. The neutrons produced can also be used to create more energy and fuel in reactions like

Equation:

$$n + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + \gamma \qquad (20.68\ \text{MeV})$$

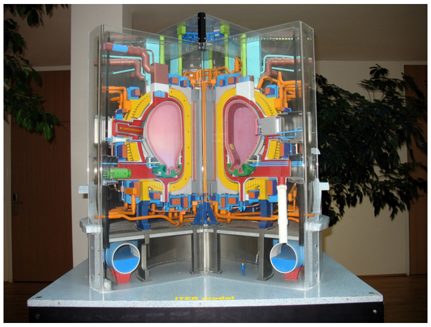
and
Equation:

$$n + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + \gamma \qquad (2.22\ \text{MeV}).$$

Note that these last two reactions, and ${}^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{4}\text{He} + \gamma$, put most of their energy output into the γ ray, and such energy is difficult to utilize.

The three keys to practical fusion energy generation are to achieve the temperatures necessary to make the reactions likely, to raise the density of the fuel, and to confine it long enough to produce large amounts of energy. These three factors—temperature, density, and time—complement one another, and so a deficiency in one can be compensated for by the others. **Ignition** is defined to occur when the reactions produce enough energy to be self-sustaining after external energy input is cut off. This goal, which must be reached before commercial plants can be a reality, has not been achieved. Another milestone, called **break-even**, occurs when the fusion power produced equals the heating power input. Break-even has nearly been reached and gives hope that ignition and commercial plants may become a reality in a few decades.

Two techniques have shown considerable promise. The first of these is called **magnetic confinement** and uses the property that charged particles have difficulty crossing magnetic field lines. The tokamak, shown in [link], has shown particular promise. The tokamak's toroidal coil confines charged particles into a circular path with a helical twist due to the circulating ions themselves. In 1995, the Tokamak Fusion Test Reactor at Princeton in the US achieved world-record plasma temperatures as high as 500 million degrees Celsius. This facility operated between 1982 and 1997. A joint international effort is underway in France to build a tokamak-type reactor that will be the stepping stone to commercial power. ITER, as it is called, will be a full-scale device that aims to demonstrate the feasibility of fusion energy. It will generate 500 MW of power for extended periods of time and will achieve break-even conditions. It will study plasmas in conditions similar to those expected in a fusion power plant. Completion is scheduled for 2018.



(a) Artist’s rendition of ITER, a tokamak-type fusion reactor being built in southern France. It is hoped that this gigantic machine will reach the break-even point. Completion is scheduled for 2018. (credit: Stephan Mosel, Flickr)

The second promising technique aims multiple lasers at tiny fuel pellets filled with a mixture of deuterium and tritium. Huge power input heats the fuel, evaporating the confining pellet and crushing the fuel to high density with the expanding hot plasma produced. This technique is called **inertial confinement**, because the fuel’s inertia prevents it from escaping before significant fusion can take place. Higher densities have been reached than with tokamaks, but with smaller confinement times. In 2009, the Lawrence Livermore Laboratory (CA) completed a laser fusion device with 192 ultraviolet laser beams that are focused upon a D-T pellet (see [link]).

National Ignition Facility (CA). This image shows a laser bay where 192 laser beams will focus onto a small D-T target, producing fusion. (credit: Lawrence Livermore National Laboratory, Lawrence Livermore National Security, LLC, and the Department of Energy)

Example:
Calculating Energy and Power from Fusion

(a) Calculate the energy released by the fusion of a 1.00-kg mixture of deuterium and tritium, which produces helium. There are equal numbers of deuterium and tritium nuclei in the mixture.

(b) If this takes place continuously over a period of a year, what is the average power output?

Strategy

According to ${}^{2}\text{H} + {}^{3}\text{H} \rightarrow {}^{4}\text{He} + n$, the energy per reaction is 17.59 MeV. To find the total energy released, we must find the number of deuterium and tritium atoms in a kilogram. Deuterium has an atomic mass of about 2 and tritium has an atomic mass of about 3, for a total of about 5 g per mole of reactants or about 200 mol in 1.00 kg. To get a more precise figure, we will use the atomic masses from Appendix A. The power output is best expressed in watts, and so the energy output needs to be calculated in joules and then divided by the number of seconds in a year.

Solution for (a)

The atomic mass of deuterium (${}^{2}\text{H}$) is 2.014102 u, while that of tritium (${}^{3}\text{H}$) is 3.016049 u, for a total of 5.032151 u per reaction. So a mole of reactants has a mass of 5.03 g, and in 1.00 kg there are $(1000\text{ g})/(5.03\text{ g/mol}){=}198.8\text{ mol of reactants}$. The number of reactions that take place is therefore

Equation:

$$(198.8\text{ mol})\left(6.02\times 10^{23}\text{ mol}^{-1}\right) = 1.20\times 10^{26}\text{ reactions.}$$

The total energy output is the number of reactions times the energy per reaction:

Equation:

$$\begin{aligned} E &= \left(1.20\times 10^{26}\text{ reactions}\right)(17.59\text{ MeV/reaction})\left(1.602\times 10^{-13}\text{ J/MeV}\right)\\ &= 3.37\times 10^{14}\text{ J.}\end{aligned}$$

Solution for (b)

Power is energy per unit time. One year has 3.16×10^{7} s, so

Equation:

$$\begin{aligned} P &= \frac{E}{t} = \frac{3.37\times 10^{14}\text{ J}}{3.16\times 10^{7}\text{ s}}\\ &= 1.07\times 10^{7}\text{ W} = 10.7\text{ MW.}\end{aligned}$$

Discussion

By now we expect nuclear processes to yield large amounts of energy, and we are not disappointed here. The energy output of 3.37×10^{14} J from fusing 1.00 kg of deuterium and tritium is equivalent to 2.6 million gallons of gasoline and about eight times the energy output of the bomb that destroyed Hiroshima. Yet the average backyard swimming pool has about 6 kg of deuterium in it, so that fuel is plentiful if it can be utilized in a controlled manner. The average power output over a year is more than 10 MW, impressive but a bit small for a commercial power plant. About 32 times this power output would allow generation of 100 MW of electricity, assuming an efficiency of one-third in converting the fusion energy to electrical energy.

Section Summary

- Nuclear fusion is a reaction in which two nuclei are combined to form a larger nucleus. It releases energy when light nuclei are fused to form medium-mass nuclei.
- Fusion is the source of energy in stars, with the proton-proton cycle,
  Equation:
  $$^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}} \qquad (0.42\text{ MeV})$$
  Equation:
  $$^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma \qquad (5.49\text{ MeV})$$
  Equation:
  $$^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H} \qquad (12.86\text{ MeV})$$
  being the principal sequence of energy-producing reactions in our Sun.
- The overall effect of the proton-proton cycle is
  Equation:
  $$2e^{-} + 4{}^{1}\text{H} \rightarrow {}^{4}\text{He} + 2v_{\text{e}} + 6\gamma \qquad (26.7\text{ MeV}),$$
  where the 26.7 MeV includes the energy of the positrons emitted and annihilated.
- Attempts to utilize controlled fusion as an energy source on Earth are related to deuterium and tritium, and the reactions play important roles.
- Ignition is the condition under which controlled fusion is self-sustaining; it has not yet been achieved. Break-even, in which the fusion energy output is as great as the external energy input, has nearly been achieved.
- Magnetic confinement and inertial confinement are the two methods being developed for heating fuel to sufficiently high temperatures, at sufficient density, and for sufficiently long times to achieve ignition. The first method uses magnetic fields

and the second method uses the momentum of impinging laser beams for confinement.

Conceptual Questions

Exercise:

Problem: Why does the fusion of light nuclei into heavier nuclei release energy?

Exercise:

Problem:

Energy input is required to fuse medium-mass nuclei, such as iron or cobalt, into more massive nuclei. Explain why.

Exercise:

Problem:

In considering potential fusion reactions, what is the advantage of the reaction $^{2}\text{H} + {}^{3}\text{H} \rightarrow {}^{4}\text{He} + n$ over the reaction $^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + n$?

Exercise:

Problem:

Give reasons justifying the contention made in the text that energy from the fusion reaction $^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{4}\text{He} + \gamma$ is relatively difficult to capture and utilize.

Problems & Exercises

Exercise:

Problem:

Verify that the total number of nucleons, total charge, and electron family number are conserved for each of the fusion reactions in the proton-proton cycle in

Equation:

$$^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}},$$

Equation:

$$^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma,$$

and

Equation:

$$^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H}.$$

(List the value of each of the conserved quantities before and after each of the reactions.)

Solution:

(a) A=1+1=2, Z=1+1=1+1, efn $= 0 = -1 + 1$

(b) A=1+2=3, Z=1+1=2, efn=0=0

(c) A=3+3=4+1+1, Z=2+2=2+1+1, efn=0=0

Exercise:

Problem:

Calculate the energy output in each of the fusion reactions in the proton-proton cycle, and verify the values given in the above summary.

Exercise:

Problem:

Show that the total energy released in the proton-proton cycle is 26.7 MeV, considering the overall effect in $^{1}\text{H} + {}^{1}\text{H} \rightarrow {}^{2}\text{H} + e^{+} + v_{\text{e}}$, $^{1}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + \gamma$, and $^{3}\text{He} + {}^{3}\text{He} \rightarrow {}^{4}\text{He} + {}^{1}\text{H} + {}^{1}\text{H}$ and being certain to include the annihilation energy.

Solution:

$$\begin{aligned} E &= (m_{\text{i}} - m_{\text{f}})c^{2} \\ &= \left[4m\left({}^{1}\text{H}\right) - m\left({}^{4}\text{He}\right)\right]c^{2} \\ &= [4(1.007825) - 4.002603](931.5\text{ MeV}) \\ &= 26.73\text{ MeV} \end{aligned}$$

Exercise:

Problem:

Verify by listing the number of nucleons, total charge, and electron family number before and after the cycle that these quantities are conserved in the overall proton-proton cycle in $2e^{-} + 4^{1}\text{H} \rightarrow {}^{4}\text{He} + 2v_{\text{e}} + 6\gamma$.

Exercise:

Problem:

The energy produced by the fusion of a 1.00-kg mixture of deuterium and tritium was found in Example Calculating Energy and Power from Fusion. Approximately how many kilograms would be required to supply the annual energy use in the United States?

Solution:

3.12×10^5 kg (about 200 tons)

Exercise:

Problem:

Tritium is naturally rare, but can be produced by the reaction $n + {}^2\text{H} \rightarrow {}^3\text{H} + \gamma$. How much energy in MeV is released in this neutron capture?

Exercise:

Problem: Two fusion reactions mentioned in the text are

$n + {}^3\text{He} \rightarrow {}^4\text{He} + \gamma$

and

$n + {}^1\text{H} \rightarrow {}^2\text{H} + \gamma$.

Both reactions release energy, but the second also creates more fuel. Confirm that the energies produced in the reactions are 20.58 and 2.22 MeV, respectively. Comment on which product nuclide is most tightly bound, ${}^4\text{He}$ or ${}^2\text{H}$.

Solution:

$$\begin{array}{lll} E & = & (m_\text{i} - m_\text{f})c^2 \\ E_1 & = & (1.008665 + 3.016030 - 4.002603)(931.5\text{ MeV}) \\ & = & 20.58\text{ MeV} \\ E_2 & = & (1.008665 + 1.007825 - 2.014102)(931.5\text{ MeV}) \\ & = & 2.224\text{ MeV} \end{array}$$

${}^4\text{He}$ is more tightly bound, since this reaction gives off more energy per nucleon.

Exercise:

Problem:

(a) Calculate the number of grams of deuterium in an 80,000-L swimming pool, given deuterium is 0.0150% of natural hydrogen.

(b) Find the energy released in joules if this deuterium is fused via the reaction $^{2}\text{H} + {}^{2}\text{H} \rightarrow {}^{3}\text{He} + n$.

(c) Could the neutrons be used to create more energy?

(d) Discuss the amount of this type of energy in a swimming pool as compared to that in, say, a gallon of gasoline, also taking into consideration that water is far more abundant.

Exercise:

Problem:

How many kilograms of water are needed to obtain the 198.8 mol of deuterium, assuming that deuterium is 0.01500% (by number) of natural hydrogen?

Solution:

1.19×10^{4} kg

Exercise:

Problem: The power output of the Sun is 4×10^{26} W.

(a) If 90% of this is supplied by the proton-proton cycle, how many protons are consumed per second?

(b) How many neutrinos per second should there be per square meter at the Earth from this process? This huge number is indicative of how rarely a neutrino interacts, since large detectors observe very few per day.

Exercise:

Problem:

Another set of reactions that result in the fusing of hydrogen into helium in the Sun and especially in hotter stars is called the carbon cycle. It is

Equation:

$$\begin{array}{lll} {}^{12}\mathrm{C} + {}^{1}\mathrm{H} & \rightarrow & {}^{13}\mathrm{N} + \gamma \\ {}^{13}\mathrm{N} & \rightarrow & {}^{13}\mathrm{C} + e^{+} + v_e \\ {}^{13}\mathrm{C} + {}^{1}\mathrm{H} & \rightarrow & {}^{14}\mathrm{N} + \gamma \\ {}^{14}\mathrm{N} + {}^{1}\mathrm{H} & \rightarrow & {}^{15}\mathrm{O} + \gamma \\ {}^{15}\mathrm{O} & \rightarrow & {}^{15}\mathrm{N} + e^{+} + v_e \\ {}^{15}\mathrm{N} + {}^{1}\mathrm{H} & \rightarrow & {}^{12}\mathrm{C} + {}^{4}\mathrm{He}. \end{array}$$

Write down the overall effect of the carbon cycle (as was done for the proton-proton cycle in $2e^{-} + 4{}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 2v_e + 6\gamma$). Note the number of protons (${}^{1}\mathrm{H}$) required and assume that the positrons (e^{+}) annihilate electrons to form more γ rays.

Solution:

$2e^{-} + 4{}^{1}\mathrm{H} \rightarrow {}^{4}\mathrm{He} + 7\gamma + 2v_e$

Exercise:

Problem:

(a) Find the total energy released in MeV in each carbon cycle (elaborated in the above problem) including the annihilation energy.

(b) How does this compare with the proton-proton cycle output?

Exercise:

Problem:

Verify that the total number of nucleons, total charge, and electron family number are conserved for each of the fusion reactions in the carbon cycle given in the above problem. (List the value of each of the conserved quantities before and after each of the reactions.)

Solution:

(a) A=12+1=13, Z=6+1=7, efn $= 0 = 0$

(b) A=13=13, Z=7=6+1, efn $= 0 = -1 + 1$

(c) A=$13 + 1$=14, Z=6+1=7, efn $= 0 = 0$

(d) A=$14 + 1$=15, Z=7+1=8, efn $= 0 = 0$

(e) A=15=15, Z=8=7+1, efn $= 0 = -1 + 1$

(f) $A{=}15+1{=}12+4$, $Z{=}7{+}1{=}6+2$, $\text{efn}=0=0$

Exercise:

Problem: Integrated Concepts

The laser system tested for inertial confinement can produce a 100-kJ pulse only 1.00 ns in duration. (a) What is the power output of the laser system during the brief pulse?

(b) How many photons are in the pulse, given their wavelength is 1.06 µm?

(c) What is the total momentum of all these photons?

(d) How does the total photon momentum compare with that of a single 1.00 MeV deuterium nucleus?

Exercise:

Problem: Integrated Concepts

Find the amount of energy given to the ^4He nucleus and to the γ ray in the reaction $n+^3\text{He}\rightarrow^4\text{He}+\gamma$, using the conservation of momentum principle and taking the reactants to be initially at rest. This should confirm the contention that most of the energy goes to the γ ray.

Solution:

$E_{\gamma}=20.6\text{ MeV}$

$E_{^4\text{He}}=5.68\times10^{-2}\text{ MeV}$

Exercise:

Problem: Integrated Concepts

(a) What temperature gas would have atoms moving fast enough to bring two ^3He nuclei into contact? Note that, because both are moving, the average kinetic energy only needs to be half the electric potential energy of these doubly charged nuclei when just in contact with one another.

(b) Does this high temperature imply practical difficulties for doing this in controlled fusion?

Exercise:

Problem: **Integrated Concepts**

(a) Estimate the years that the deuterium fuel in the oceans could supply the energy needs of the world. Assume world energy consumption to be ten times that of the United States which is 8×10^{19} J/y and that the deuterium in the oceans could be converted to energy with an efficiency of 32%. You must estimate or look up the amount of water in the oceans and take the deuterium content to be 0.015% of natural hydrogen to find the mass of deuterium available. Note that approximate energy yield of deuterium is 3.37×10^{14} J/kg.

(b) Comment on how much time this is by any human measure. (It is not an unreasonable result, only an impressive one.)

Solution:

(a) 3×10^{9} y

(b) This is approximately half the lifetime of the Earth.

Glossary

break-even
when fusion power produced equals the heating power input

ignition
when a fusion reaction produces enough energy to be self-sustaining after external energy input is cut off

inertial confinement
a technique that aims multiple lasers at tiny fuel pellets evaporating and crushing them to high density

magnetic confinement
a technique in which charged particles are trapped in a small region because of difficulty in crossing magnetic field lines

nuclear fusion
a reaction in which two nuclei are combined, or fused, to form a larger nucleus

proton-proton cycle
the combined reactions $^{1}\text{H}+^{1}\text{H}\rightarrow^{2}\text{H}+e^{+}+v_{\text{e}}$, $^{1}\text{H}+^{2}\text{H}\rightarrow^{3}\text{He}+\gamma$, and $^{3}\text{He}+^{3}\text{He}\rightarrow^{4}\text{He}+^{1}\text{H}+^{1}\text{H}$

Fission

- Define nuclear fission.
- Discuss how fission fuel reacts and describe what it produces.
- Describe controlled and uncontrolled chain reactions.

Nuclear fission is a reaction in which a nucleus is split (or *fissured*). Controlled fission is a reality, whereas controlled fusion is a hope for the future. Hundreds of nuclear fission power plants around the world attest to the fact that controlled fission is practical and, at least in the short term, economical, as seen in [link]. Whereas nuclear power was of little interest for decades following TMI and Chernobyl (and now Fukushima Daiichi), growing concerns over global warming has brought nuclear power back on the table as a viable energy alternative. By the end of 2009, there were 442 reactors operating in 30 countries, providing 15% of the world's electricity. France provides over 75% of its electricity with nuclear power, while the US has 104 operating reactors providing 20% of its electricity. Australia and New Zealand have none. China is building nuclear power plants at the rate of one start every month.

The people living near this nuclear power plant have no measurable exposure to radiation that is traceable to the plant. About 16% of the world's electrical power is generated by controlled nuclear fission in such plants. The cooling towers are the most prominent features but are not unique to nuclear power. The reactor is in the small domed building to the left of the towers. (credit: Kalmthouts)

Fission is the opposite of fusion and releases energy only when heavy nuclei are split. As noted in Fusion, energy is released if the products of a nuclear reaction have a greater binding energy per nucleon (BE/A) than the parent nuclei. [link] shows that BE/A is greater for medium-mass nuclei than heavy nuclei, implying that when a heavy nucleus is split, the products have less mass per nucleon, so that mass is destroyed and energy is released in the reaction. The amount of energy per fission reaction can be large, even by nuclear standards. The graph in [link] shows BE/A to be about 7.6 MeV/nucleon for the heaviest nuclei (A about 240), while BE/A is about 8.6 MeV/nucleon for nuclei having A about 120. Thus, if a heavy nucleus splits in half, then about 1 MeV per nucleon, or approximately 240 MeV per fission, is released. This is about 10 times the energy per fusion reaction, and about 100 times the energy of the average α, β, or γ decay.

Example:
Calculating Energy Released by Fission
Calculate the energy released in the following spontaneous fission reaction:
Equation:

$$^{238}\text{U} \rightarrow {}^{95}\text{Sr} + {}^{140}\text{Xe} + 3n$$

given the atomic masses to be $m(^{238}\text{U}) = 238.050784\text{ u}$, $m(^{95}\text{Sr}) = 94.919388\text{ u}$, $m(^{140}\text{Xe}) = 139.921610\text{ u}$, and $m(n) = 1.008665\text{ u}$.

Strategy
As always, the energy released is equal to the mass destroyed times c^2 , so we must find the difference in mass between the parent ^{238}U and the fission products.

Solution
The products have a total mass of
Equation:

$$\begin{aligned} m_{\text{products}} &= 94.919388\text{ u} + 139.921610\text{ u} + 3(1.008665\text{ u}) \\ &= 237.866993\text{ u}. \end{aligned}$$

The mass lost is the mass of ^{238}U minus m_{products}, or
Equation:

$$\Delta m = 238.050784\text{ u} - 237.8669933\text{ u} = 0.183791\text{ u},$$

so the energy released is
Equation:

$$\begin{aligned} E &= (\Delta m)c^2 \\ &= (0.183791\text{ u})\frac{931.5\text{ MeV}/c^2}{\text{u}}c^2 = 171.2\text{ MeV}. \end{aligned}$$

Discussion
A number of important things arise in this example. The 171-MeV energy released is large, but a little less than the earlier estimated 240 MeV. This is because this fission reaction produces neutrons and does not split the nucleus into two equal parts. Fission of a given nuclide, such as ^{238}U , does not always produce the same products. Fission is a statistical process in which an entire range of products are produced with various probabilities. Most fission produces neutrons, although the number varies with each fission. This is an extremely important aspect of fission, because *neutrons can induce more fission*, enabling self-sustaining chain reactions.

Spontaneous fission can occur, but this is usually not the most common decay mode for a given nuclide. For example, ^{238}U can spontaneously fission, but it decays mostly by α emission. Neutron-induced fission is crucial as seen in [link]. Being chargeless, even low-energy neutrons can strike a nucleus and be absorbed once they feel the attractive nuclear force. Large nuclei are described by a **liquid drop model** with surface tension and oscillation modes, because the large number of nucleons act like atoms in a drop. The neutron is attracted and thus, deposits energy, causing the nucleus to deform as a liquid drop. If stretched enough, the nucleus narrows in the middle. The number of nucleons in contact and the strength of the nuclear force binding the nucleus together are reduced. Coulomb repulsion between the two ends then succeeds in fissioning the nucleus, which pops like a water drop into two large pieces and a few neutrons. **Neutron-induced fission** can be written as

Equation:

$$n + {}^{A}\mathrm{X} \to \mathrm{FF}_1 + \mathrm{FF}_2 + \mathrm{xn},$$

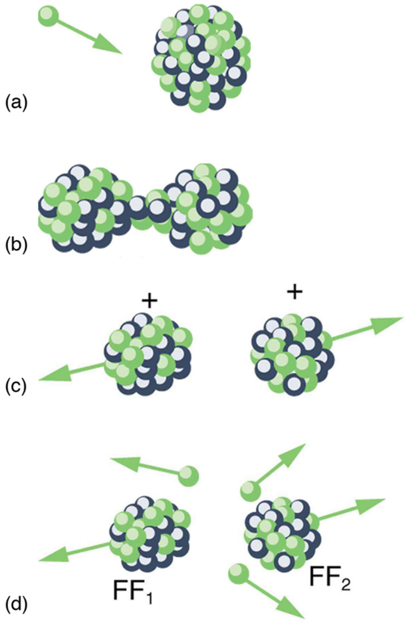
where FF_1 and FF_2 are the two daughter nuclei, called **fission fragments**, and x is the number of neutrons produced. Most often, the masses of the fission fragments are not the same. Most of the released energy goes into the kinetic energy of the fission fragments, with the remainder going into the neutrons and excited states of the fragments. Since neutrons can induce fission, a self-sustaining chain reaction is possible, provided more than one neutron is produced on average — that is, if $x > 1$ in $n + {}^{A}\mathrm{X} \to \mathrm{FF}_1 + \mathrm{FF}_2 + \mathrm{xn}$. This can also be seen in [link].

An example of a typical neutron-induced fission reaction is

Equation:

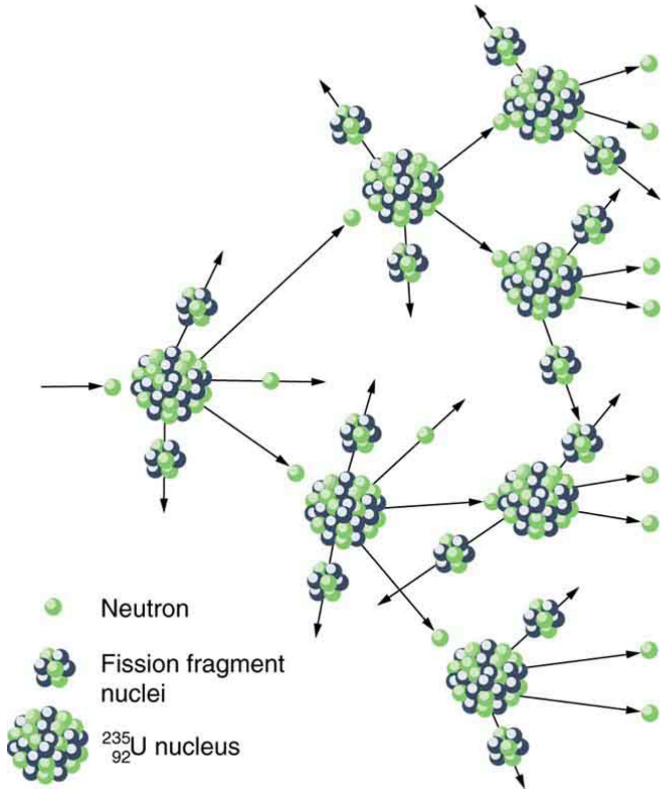
$$n + {}^{235}_{92}\mathrm{U} \to {}^{142}_{56}\mathrm{Ba} + {}^{91}_{36}\mathrm{Kr} + 3n.$$

Note that in this equation, the total charge remains the same (is conserved): $92 + 0 = 56 + 36$. Also, as far as whole numbers are concerned, the mass is constant: $1 + 235 = 142 + 91 + 3$. This is not true when we consider the masses out to 6 or 7 significant places, as in the previous example.



Neutron-induced

fission is shown. First, energy is put into this large nucleus when it absorbs a neutron. Acting like a struck liquid drop, the nucleus deforms and begins to narrow in the middle. Since fewer nucleons are in contact, the repulsive Coulomb force is able to break the nucleus into two parts with some neutrons also flying away.

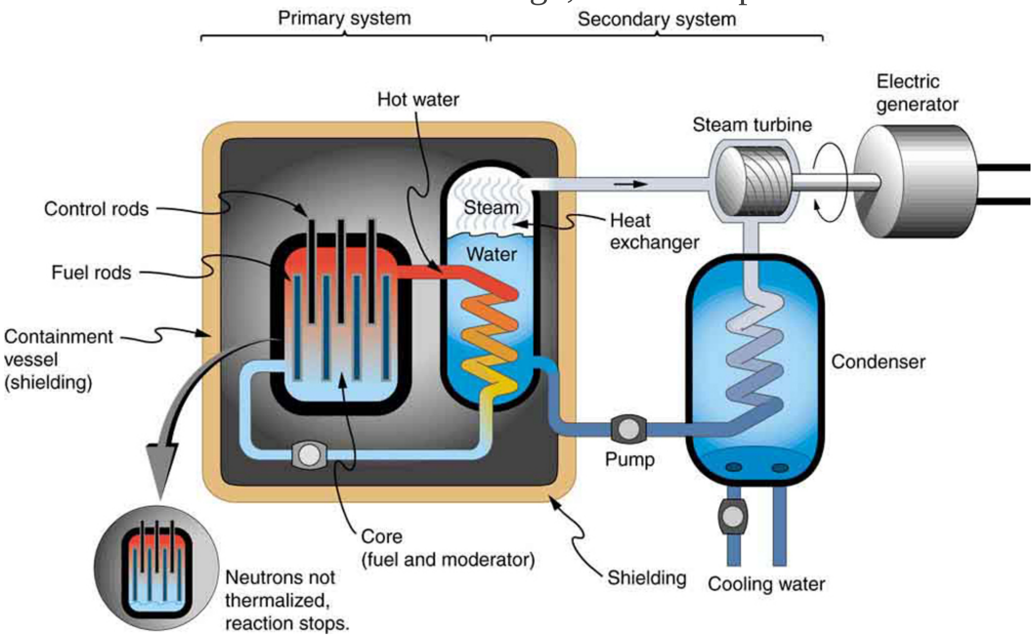


A chain reaction can produce self-sustained fission if each fission produces enough neutrons to induce at least one more fission. This depends on several factors, including how many neutrons are produced in an average fission and how easy it is to make a particular type of nuclide fission.

Not every neutron produced by fission induces fission. Some neutrons escape the fissionable material, while others interact with a nucleus without making it fission. We can enhance the number of fissions produced by neutrons by having a large amount of fissionable material. The minimum amount necessary for self-sustained fission of a given nuclide is called its **critical mass**. Some nuclides, such as $^{239}\mathrm{Pu}$, produce more neutrons per fission than others, such as $^{235}\mathrm{U}$. Additionally, some nuclides are easier to make fission than others. In particular, $^{235}\mathrm{U}$ and $^{239}\mathrm{Pu}$ are easier to fission than the much more abundant $^{238}\mathrm{U}$. Both factors affect critical mass, which is smallest for $^{239}\mathrm{Pu}$.

The reason $^{235}\mathrm{U}$ and $^{239}\mathrm{Pu}$ are easier to fission than $^{238}\mathrm{U}$ is that the nuclear force is more attractive for an even number of neutrons in a nucleus than for an odd number. Consider that $^{235}_{92}\mathrm{U}_{143}$ has 143 neutrons, and $^{239}_{94}\mathrm{P}_{145}$ has 145 neutrons, whereas $^{238}_{92}\mathrm{U}_{146}$ has 146. When a neutron encounters a nucleus with an odd number of neutrons, the nuclear force is more attractive, because the additional neutron will make the number even. About 2-MeV more energy is deposited in the resulting nucleus than would be the case if the number of neutrons was already even. This extra energy produces greater deformation, making fission more likely. Thus, $^{235}\mathrm{U}$ and $^{239}\mathrm{Pu}$ are superior fission fuels. The isotope $^{235}\mathrm{U}$ is only 0.72 % of natural uranium, while $^{238}\mathrm{U}$ is 99.27%, and $^{239}\mathrm{Pu}$ does not exist in nature. Australia has the largest deposits of uranium in the world, standing at 28% of the total. This is followed by Kazakhstan and Canada. The US has only 3% of global reserves.

Most fission reactors utilize $^{235}\mathrm{U}$, which is separated from $^{238}\mathrm{U}$ at some expense. This is called enrichment. The most common separation method is gaseous diffusion of uranium hexafluoride (UF_6) through membranes. Since $^{235}\mathrm{U}$ has less mass than $^{238}\mathrm{U}$, its UF_6 molecules have higher average velocity at the same temperature and diffuse faster. Another interesting characteristic of $^{235}\mathrm{U}$ is that it preferentially absorbs very slow moving neutrons (with energies a fraction of an eV), whereas fission reactions produce fast neutrons with energies in the order of an MeV. To make a self-sustained fission reactor with ${}^{235}\text{U}$, it is thus necessary to slow down (“thermalize”) the neutrons. Water is very effective, since neutrons collide with protons in water molecules and lose energy. [link] shows a schematic of a reactor design, called the pressurized water reactor.



A pressurized water reactor is cleverly designed to control the fission of large amounts of ${}^{235}\text{U}$, while using the heat produced in the fission reaction to create steam for generating electrical energy. Control rods adjust neutron flux so that criticality is obtained, but not exceeded. In case the reactor overheats and boils the water away, the chain reaction terminates, because water is needed to thermalize the neutrons. This inherent safety feature can be overwhelmed in extreme circumstances.

Control rods containing nuclides that very strongly absorb neutrons are used to adjust neutron flux. To produce large power, reactors contain hundreds to thousands of critical masses, and the chain reaction easily becomes self-sustaining, a condition called **criticality**. Neutron flux should be carefully regulated to avoid an exponential increase in fissions, a condition called **supercriticality**. Control rods help prevent overheating, perhaps even a meltdown or explosive disassembly. The water that is used to thermalize neutrons, necessary to get them to induce fission in ^{235}U , and achieve criticality, provides a negative feedback for temperature increases. In case the reactor overheats and boils the water to steam or is breached, the absence of water kills the chain reaction. Considerable heat, however, can still be generated by the reactor's radioactive fission products. Other safety features, thus, need to be incorporated in the event of a *loss of coolant* accident, including auxiliary cooling water and pumps.

Example:
Calculating Energy from a Kilogram of Fissionable Fuel
Calculate the amount of energy produced by the fission of 1.00 kg of ^{235}U , given the average fission reaction of ^{235}U produces 200 MeV.
Strategy
The total energy produced is the number of ^{235}U atoms times the given energy per ^{235}U fission. We should therefore find the number of ^{235}U atoms in 1.00 kg.
Solution
The number of ^{235}U atoms in 1.00 kg is Avogadro's number times the number of moles. One mole of ^{235}U has a mass of 235.04 g; thus, there are $(1000\text{ g})/(235.04\text{ g/mol}) = 4.25\text{ mol}$. The number of ^{235}U atoms is therefore,
Equation:

$$(4.25\text{ mol})\left(6.02\times10^{23}\ ^{235}\text{U/mol}\right) = 2.56\times10^{24}\ ^{235}\text{U}.$$

So the total energy released is
Equation:

$$\begin{array}{lll} E & = & \left(2.56\times10^{24}\ ^{235}\text{U}\right)\left(\frac{200\text{ MeV}}{^{235}\text{U}}\right)\left(\frac{1.60\times10^{-13}\text{ J}}{\text{MeV}}\right) \\ & = & 8.21\times10^{13}\text{ J}. \end{array}$$

Discussion
This is another impressively large amount of energy, equivalent to about 14,000 barrels of crude oil or 600,000 gallons of gasoline. But, it is only one-fourth the energy produced by the fusion of a kilogram mixture of deuterium and tritium as seen in [link]. Even though each fission reaction yields about ten times the energy of a fusion reaction, the energy per kilogram of fission fuel is less, because there are far fewer moles per kilogram of the heavy nuclides. Fission

fuel is also much more scarce than fusion fuel, and less than 1% of uranium $(\text{the }^{235}\text{U})$ is readily usable.

One nuclide already mentioned is ^{239}Pu , which has a 24,120-y half-life and does not exist in nature. Plutonium-239 is manufactured from ^{238}U in reactors, and it provides an opportunity to utilize the other 99% of natural uranium as an energy source. The following reaction sequence, called **breeding**, produces ^{239}Pu . Breeding begins with neutron capture by ^{238}U :

Equation:

$$^{238}\text{U}+n\rightarrow\,^{239}\text{U}+\gamma.$$

Uranium-239 then β^{-} decays:

Equation:

$$^{239}\text{U}\rightarrow\,^{239}\text{Np}+\beta^{-}+v_{e}(\text{t}_{1/2}=23\text{ min}).$$

Neptunium-239 also β^{-} decays:

Equation:

$$^{239}\text{Np}\rightarrow\,^{239}\text{Pu}+\beta^{-}+v_{e}(\text{t}_{1/2}=2.4\text{ d}).$$

Plutonium-239 builds up in reactor fuel at a rate that depends on the probability of neutron capture by ^{238}U (all reactor fuel contains more ^{238}U than ^{235}U). Reactors designed specifically to make plutonium are called **breeder reactors**. They seem to be inherently more hazardous than conventional reactors, but it remains unknown whether their hazards can be made economically acceptable. The four reactors at Chernobyl, including the one that was destroyed, were built to breed plutonium and produce electricity. These reactors had a design that was significantly different from the pressurized water reactor illustrated above.

Plutonium-239 has advantages over ^{235}U as a reactor fuel — it produces more neutrons per fission on average, and it is easier for a thermal neutron to cause it to fission. It is also chemically different from uranium, so it is inherently easier to separate from uranium ore. This means ^{239}Pu has a particularly small critical mass, an advantage for nuclear weapons.

Note:
PhET Explorations: Nuclear Fission
Start a chain reaction, or introduce non-radioactive isotopes to prevent one. Control energy production in a nuclear reactor!

https://archive.cnx.org/specials/01caf0d0-116f-11e6-b891-abfdaa77b03b/nuclear-fission/#sim-one-nucleus

Section Summary

- Nuclear fission is a reaction in which a nucleus is split.
- Fission releases energy when heavy nuclei are split into medium-mass nuclei.
- Self-sustained fission is possible, because neutron-induced fission also produces neutrons that can induce other fissions, $n + {}^{A}X \rightarrow \text{FF}_1 + \text{FF}_2 + \text{xn}$, where FF_1 and FF_2 are the two daughter nuclei, or fission fragments, and *x* is the number of neutrons produced.
- A minimum mass, called the critical mass, should be present to achieve criticality.
- More than a critical mass can produce supercriticality.
- The production of new or different isotopes (especially ${}^{239}\text{Pu}$) by nuclear transformation is called breeding, and reactors designed for this purpose are called breeder reactors.

Conceptual Questions

Exercise:

Problem:

Explain why the fission of heavy nuclei releases energy. Similarly, why is it that energy input is required to fission light nuclei?

Exercise:

Problem:

Explain, in terms of conservation of momentum and energy, why collisions of neutrons with protons will thermalize neutrons better than collisions with oxygen.

Exercise:

Problem:

The ruins of the Chernobyl reactor are enclosed in a huge concrete structure built around it after the accident. Some rain penetrates the building in winter, and radioactivity from the building increases. What does this imply is happening inside?

Exercise:

Problem:

Since the uranium or plutonium nucleus fissions into several fission fragments whose mass distribution covers a wide range of pieces, would you expect more residual radioactivity from fission than fusion? Explain.

Exercise:

Problem:

The core of a nuclear reactor generates a large amount of thermal energy from the decay of fission products, even when the power-producing fission chain reaction is turned off. Would this residual heat be greatest after the reactor has run for a long time or short time? What if the reactor has been shut down for months?

Exercise:

Problem:

How can a nuclear reactor contain many critical masses and not go supercritical? What methods are used to control the fission in the reactor?

Exercise:

Problem:

Why can heavy nuclei with odd numbers of neutrons be induced to fission with thermal neutrons, whereas those with even numbers of neutrons require more energy input to induce fission?

Exercise:

Problem:

Why is a conventional fission nuclear reactor not able to explode as a bomb?

Problem Exercises

Exercise:

Problem:

(a) Calculate the energy released in the neutron-induced fission (similar to the spontaneous fission in [link])

Equation:

$$n + {}^{238}\text{U} \rightarrow {}^{96}\text{Sr} + {}^{140}\text{Xe} + 3n,$$

given $m({}^{96}\text{Sr}) = 95.921750$ u and $m({}^{140}\text{Xe}) = 139.92164$. (b) This result is about 6 MeV greater than the result for spontaneous fission. Why? (c) Confirm that the total number of nucleons and total charge are conserved in this reaction.

Solution:

(a) 177.1 MeV

(b) Because the gain of an external neutron yields about 6 MeV, which is the average BE/A for heavy nuclei.

(c)
$A = 1 + 238 = 96 + 140 + 1 + 1 + 1$, $Z = 92 = 38 + 53$, $\text{efn} = 0 = 0$

Exercise:

Problem:

(a) Calculate the energy released in the neutron-induced fission reaction
Equation:

$$n+{}^{235}\text{U}\rightarrow{}^{92}\text{Kr}+{}^{142}\text{Ba}+2n,$$

given $m({}^{92}\text{Kr})=91.926269\text{ u}$ and $m({}^{142}\text{Ba})=141.916361\text{ u}.$

(b) Confirm that the total number of nucleons and total charge are conserved in this reaction.

Exercise:

Problem:

(a) Calculate the energy released in the neutron-induced fission reaction
Equation:

$$n+{}^{239}\text{Pu}\rightarrow{}^{96}\text{Sr}+{}^{140}\text{Ba}+4n,$$

given $m({}^{96}\text{Sr})=95.921750\text{ u}$ and $m({}^{140}\text{Ba})=139.910581\text{ u}.$

(b) Confirm that the total number of nucleons and total charge are conserved in this reaction.

Solution:

(a) 180.6 MeV

(b)
$A=1+239=96+140+1+1+1+1,\ Z=94=38+56,\ \text{efn}=0=0$

Exercise:

Problem:

Confirm that each of the reactions listed for plutonium breeding just following [link] conserves the total number of nucleons, the total charge, and electron family number.

Exercise:

Problem:

Breeding plutonium produces energy even before any plutonium is fissioned. (The primary purpose of the four nuclear reactors at Chernobyl was breeding plutonium for weapons. Electrical power was a by-product used by the civilian population.) Calculate the energy produced in each of the reactions listed for plutonium breeding just following [link]. The pertinent masses are $m(^{239}\text{U}) = 239.054289\text{ u}$, $m(^{239}\text{Np}) = 239.052932\text{ u}$, and $m(^{239}\text{Pu}) = 239.052157\text{ u}$.

Solution:

$^{238}\text{U} + n \ \rightarrow\ ^{239}\text{U} + \gamma$ 4.81 MeV

$^{239}\text{U} \rightarrow\ ^{239}\text{Np} + \beta^{-} + v_e$ 0.753 MeV

$^{239}\text{Np} \rightarrow\ ^{239}\text{Pu} + \beta^{-} + v_e$ 0.211 MeV

Exercise:

Problem:

The naturally occurring radioactive isotope ^{232}Th does not make good fission fuel, because it has an even number of neutrons; however, it can be bred into a suitable fuel (much as ^{238}U is bred into ^{239}P).

(a) What are Z and N for ^{232}Th?

(b) Write the reaction equation for neutron captured by ^{232}Th and identify the nuclide ^{A}X produced in $n +\ ^{232}\text{Th} \rightarrow\ ^{A}X + \gamma$.

(c) The product nucleus β^{-} decays, as does its daughter. Write the decay equations for each, and identify the final nucleus.

(d) Confirm that the final nucleus has an odd number of neutrons, making it a better fission fuel.

(e) Look up the half-life of the final nucleus to see if it lives long enough to be a useful fuel.

Exercise:

Problem:

The electrical power output of a large nuclear reactor facility is 900 MW. It has a 35.0% efficiency in converting nuclear power to electrical.

(a) What is the thermal nuclear power output in megawatts?

(b) How many ^{235}U nuclei fission each second, assuming the average fission produces 200 MeV?

(c) What mass of ^{235}U is fissioned in one year of full-power operation?

Solution:

(a) $2.57 \times 10^3 \text{ MW}$

(b) $8.03 \times 10^{19} \text{ fission/s}$

(c) 991 kg

Exercise:

Problem:

A large power reactor that has been in operation for some months is turned off, but residual activity in the core still produces 150 MW of power. If the average energy per decay of the fission products is 1.00 MeV, what is the core activity in curies?

Glossary

breeder reactors
  reactors that are designed specifically to make plutonium

breeding
  reaction process that produces ^{239}Pu

criticality
  condition in which a chain reaction easily becomes self-sustaining

critical mass

minimum amount necessary for self-sustained fission of a given nuclide

fission fragments
a daughter nuclei

liquid drop model
a model of nucleus (only to understand some of its features) in which nucleons in a nucleus act like atoms in a drop

nuclear fission
reaction in which a nucleus splits

neutron-induced fission
fission that is initiated after the absorption of neutron

supercriticality
an exponential increase in fissions

Nuclear Weapons

- Discuss different types of fission and thermonuclear bombs.
- Explain the ill effects of nuclear explosion.

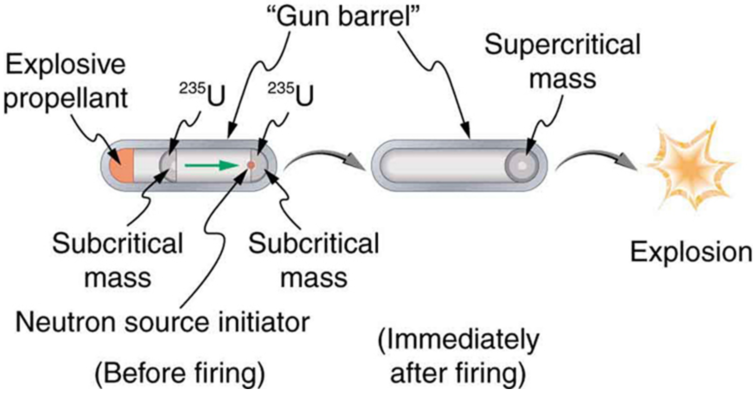
The world was in turmoil when fission was discovered in 1938. The discovery of fission, made by two German physicists, Otto Hahn and Fritz Strassman, was quickly verified by two Jewish refugees from Nazi Germany, Lise Meitner and her nephew Otto Frisch. Fermi, among others, soon found that not only did neutrons induce fission; more neutrons were produced during fission. The possibility of a self-sustained chain reaction was immediately recognized by leading scientists the world over. The enormous energy known to be in nuclei, but considered inaccessible, now seemed to be available on a large scale.

Within months after the announcement of the discovery of fission, Adolf Hitler banned the export of uranium from newly occupied Czechoslovakia. It seemed that the military value of uranium had been recognized in Nazi Germany, and that a serious effort to build a nuclear bomb had begun.

Alarmed scientists, many of them who fled Nazi Germany, decided to take action. None was more famous or revered than Einstein. It was felt that his help was needed to get the American government to make a serious effort at nuclear weapons as a matter of survival. Leo Szilard, an escaped Hungarian physicist, took a draft of a letter to Einstein, who, although pacifistic, signed the final version. The letter was for President Franklin Roosevelt, warning of the German potential to build extremely powerful bombs of a new type. It was sent in August of 1939, just before the German invasion of Poland that marked the start of World War II.

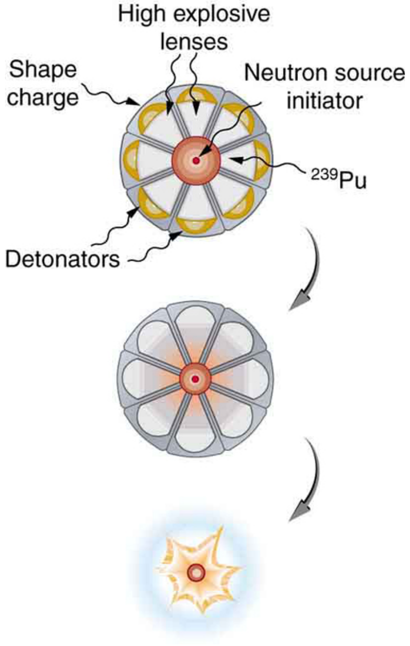
It was not until December 6, 1941, the day before the Japanese attack on Pearl Harbor, that the United States made a massive commitment to building a nuclear bomb. The top secret Manhattan Project was a crash program aimed at beating the Germans. It was carried out in remote locations, such as Los Alamos, New Mexico, whenever possible, and eventually came to cost billions of dollars and employ the efforts of more than 100,000 people. J. Robert Oppenheimer (1904–1967), whose talent and ambitions made him ideal, was chosen to head the project. The first major step was made by Enrico Fermi and his group in December 1942, when they achieved the first self-sustained nuclear reactor. This first “atomic pile”, built in a squash court at the University of Chicago, used carbon blocks to thermalize neutrons. It not only proved that the chain reaction was possible, it began the era of nuclear reactors. Glenn Seaborg, an American chemist and physicist, received the Nobel Prize in physics in 1951 for discovery of several transuranic elements, including plutonium. Carbon-moderated reactors are relatively inexpensive and simple in design and are still used for breeding plutonium, such as at Chernobyl, where two such reactors remain in operation.

Plutonium was recognized as easier to fission with neutrons and, hence, a superior fission material very early in the Manhattan Project. Plutonium availability was uncertain, and so a uranium bomb was developed simultaneously. [link] shows a gun-type bomb, which takes two subcritical uranium masses and blows them together. To get an appreciable yield, the critical mass must be held together by the explosive charges inside the cannon barrel for a few microseconds. Since the buildup of the uranium chain reaction is relatively slow, the device to hold the critical mass together can be relatively simple. Owing to the fact that the rate of spontaneous fission is low, a neutron source is triggered at the same time the critical mass is assembled.



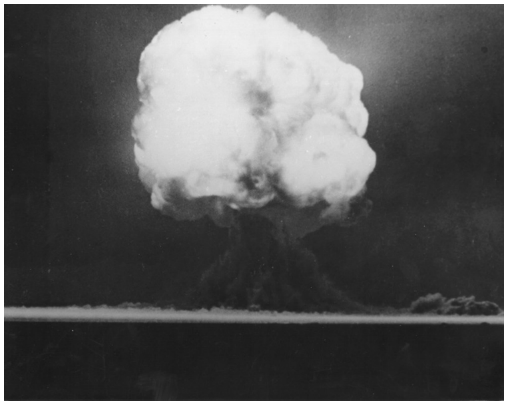
A gun-type fission bomb for ^{235}U utilizes two subcritical masses forced together by explosive charges inside a cannon barrel. The energy yield depends on the amount of uranium and the time it can be held together before it disassembles itself.

Plutonium’s special properties necessitated a more sophisticated critical mass assembly, shown schematically in [link]. A spherical mass of plutonium is surrounded by shape charges (high explosives that release most of their blast in one direction) that implode the plutonium, crushing it into a smaller volume to form a critical mass. The implosion technique is faster and more effective, because it compresses three-dimensionally rather than one-dimensionally as in the gun-type bomb. Again, a neutron source must be triggered at just the correct time to initiate the chain reaction.



An implosion created by high explosives compresses a sphere of ^{239}Pu into a critical mass. The superior fissionability of plutonium has made it the universal bomb material.

Owing to its complexity, the plutonium bomb needed to be tested before there could be any attempt to use it. On July 16, 1945, the test named Trinity was conducted in the isolated Alamogordo Desert about 200 miles south of Los Alamos (see [link]). A new age had begun. The yield of this device was about 10 kilotons (kT), the equivalent of 5000 of the largest conventional bombs.



Trinity test (1945), the first nuclear bomb (credit: United States Department of Energy)

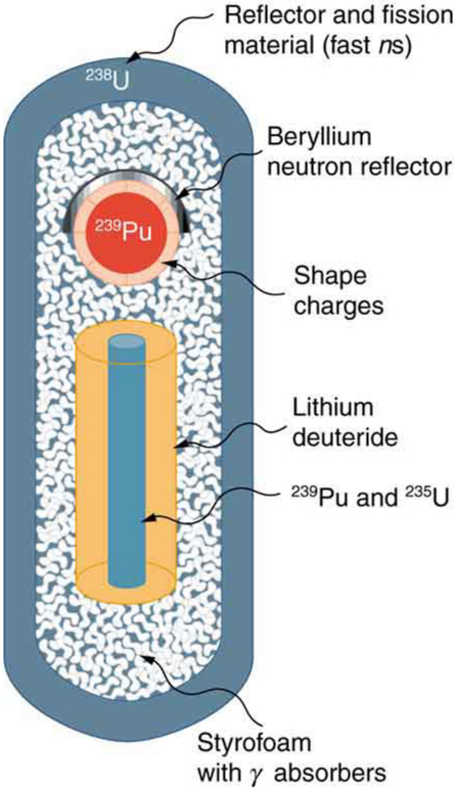
Although Germany surrendered on May 7, 1945, Japan had been steadfastly refusing to surrender for many months, forcing large casualties. Invasion plans by the Allies estimated a million casualties of their own and untold losses of Japanese lives. The bomb was viewed as a way to end the war. The first was a uranium bomb dropped on Hiroshima on August 6. Its yield of about 15 kT destroyed the city and killed an estimated 80,000 people, with 100,000 more being seriously injured (see [link]). The second was a plutonium bomb dropped on Nagasaki only three days later, on August 9. Its 20 kT yield killed at least 50,000 people, something less than Hiroshima because of the hilly terrain and the fact that it was a few kilometers off target. The Japanese were told that one bomb a week would be dropped until they surrendered unconditionally, which they did on August 14. In actuality, the United States had only enough plutonium for one more and as yet unassembled bomb.

Destruction in Hiroshima (credit: United States Federal Government)

Knowing that fusion produces several times more energy per kilogram of fuel than fission, some scientists pushed the idea of a fusion bomb starting very early on. Calling this bomb the Super, they realized that it could have another advantage over fission—high-energy neutrons would aid fusion, while they are ineffective in $^{239}\mathrm{Pu}$ fission. Thus the fusion bomb could be virtually unlimited in energy release. The first such bomb was detonated by the United States on October 31, 1952, at Eniwetok Atoll with a yield of 10 megatons (MT), about 670 times that of the fission bomb that destroyed Hiroshima. The Soviets followed with a fusion device of their own in August 1953, and a weapons race, beyond the aim of this text to discuss, continued until the end of the Cold War.

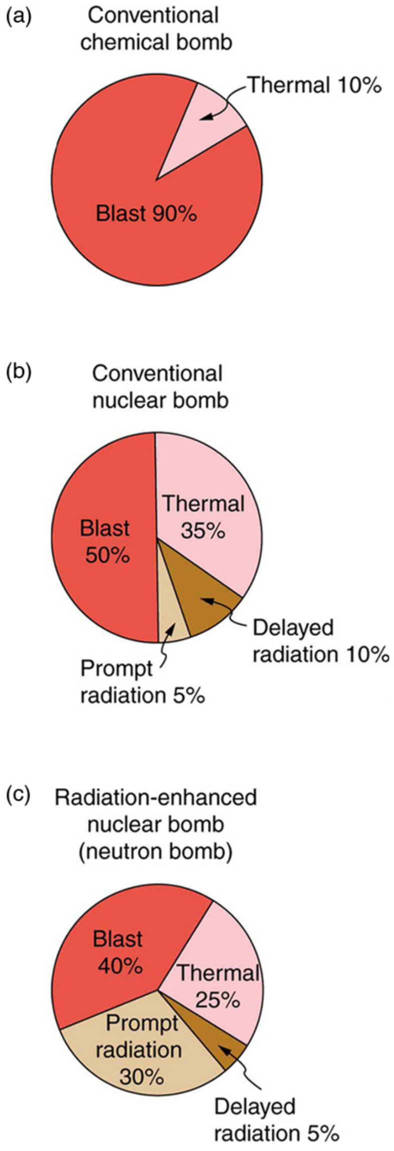
[link] shows a simple diagram of how a thermonuclear bomb is constructed. A fission bomb is exploded next to fusion fuel in the solid form of lithium deuteride. Before the shock wave blows it apart, γ rays heat and compress the fuel, and neutrons create tritium through the reaction $n+^{6}\text{Li}\rightarrow^{3}\text{H}+^{4}\text{He}$. Additional fusion and fission fuels are enclosed in a dense shell of $^{238}\mathrm{U}$. The shell reflects some of the neutrons back into the fuel to enhance its fusion, but at high internal temperatures fast neutrons are

created that also cause the plentiful and inexpensive ^{238}U to fission, part of what allows thermonuclear bombs to be so large.


Reflector and fission material (fast *n*s)
^{238}U
Beryllium neutron reflector
^{239}Pu
Shape charges
Lithium deuteride
^{239}Pu and ^{235}U
Styrofoam with γ absorbers

This schematic of a fusion bomb (H-bomb) gives some idea of how the ^{239}Pu fission trigger is used to ignite fusion fuel. Neutrons and γ rays transmit energy to the fusion fuel, create tritium from deuterium, and heat and compress the fusion fuel. The outer shell of ^{238}U serves to reflect some neutrons back into the fuel, causing more fusion, and it boosts the energy output by fissioning itself when neutron energies become high enough.

The energy yield and the types of energy produced by nuclear bombs can be varied. Energy yields in current arsenals range from about 0.1 kT to 20 MT, although the Soviets once detonated a 67 MT device. Nuclear bombs differ from conventional explosives in more than size. [link] shows the approximate fraction of energy output in various forms for conventional explosives and for two types of nuclear bombs. Nuclear bombs put a much larger fraction of their output into thermal energy than do conventional bombs, which tend to concentrate the energy in blast. Another difference is the immediate and residual radiation energy from nuclear weapons. This can be adjusted to put more energy into radiation (the so-called neutron bomb) so that the bomb can be used to irradiate advancing troops without killing friendly troops with blast and heat.



Approximate fractions of energy output by conventional and two types of nuclear weapons. In addition to yielding more energy than conventional weapons, nuclear bombs put a much larger fraction into thermal energy. This can be adjusted to enhance the radiation output to be more effective against troops. An enhanced radiation bomb is also called a neutron bomb.

At its peak in 1986, the combined arsenals of the United States and the Soviet Union totaled about 60,000 nuclear warheads. In addition, the British, French, and Chinese each have several hundred bombs of various sizes, and a few other countries have a small number. Nuclear weapons are generally divided into two categories. Strategic nuclear weapons are those intended for military targets, such as bases and missile complexes, and moderate to large cities. There were about 20,000 strategic weapons in 1988. Tactical weapons are intended for use in smaller battles. Since the collapse of the Soviet Union and the end of the Cold War in 1989, most of the 32,000 tactical weapons (including Cruise missiles, artillery shells, land mines, torpedoes, depth charges, and backpacks) have been demobilized, and parts of the strategic weapon systems are being dismantled with warheads and missiles being disassembled. According to the Treaty of Moscow of 2002, Russia and the United States have been required to reduce their strategic nuclear arsenal down to about 2000 warheads each.

A few small countries have built or are capable of building nuclear bombs, as are some terrorist groups. Two things are needed—a minimum level of technical expertise and sufficient fissionable material. The first is easy. Fissionable material is controlled but is also available. There are international agreements and organizations that attempt to control nuclear proliferation, but it is increasingly difficult given the availability of fissionable material and the small amount needed for a crude bomb. The production of fissionable fuel itself is technologically difficult. However, the presence of large amounts of such material worldwide, though in the hands of a few, makes control and accountability crucial.

Section Summary

- There are two types of nuclear weapons—fission bombs use fission alone, whereas thermonuclear bombs use fission to ignite fusion.
- Both types of weapons produce huge numbers of nuclear reactions in a very short time.
- Energy yields are measured in kilotons or megatons of equivalent conventional explosives and range from 0.1 kT to more than 20 MT.
- Nuclear bombs are characterized by far more thermal output and nuclear radiation output than conventional explosives.

Conceptual Questions

Exercise:

Problem:

What are some of the reasons that plutonium rather than uranium is used in all fission bombs and as the trigger in all fusion bombs?

Exercise:

Problem:

Use the laws of conservation of momentum and energy to explain how a shape charge can direct most of the energy released in an explosion in a specific direction. (Note that this is similar to the situation in guns and cannons—most of the energy goes into the bullet.)

Exercise:

Problem:

How does the lithium deuteride in the thermonuclear bomb shown in [link] supply tritium $\left({}^{3}\text{H}\right)$ as well as deuterium $\left({}^{2}\text{H}\right)$?

Exercise:

Problem:

Fallout from nuclear weapons tests in the atmosphere is mainly ^{90}Sr and ^{137}Cs, which have 28.6- and 32.2-y half-lives, respectively. Atmospheric tests were terminated in most countries in 1963, although China only did so in 1980. It has been found that environmental activities of these two isotopes are decreasing faster than their half-lives. Why might this be?

Problems & Exercises

Exercise:

Problem: Find the mass converted into energy by a 12.0-kT bomb.

Solution:

0.56 g

Exercise:

Problem: What mass is converted into energy by a 1.00-MT bomb?

Exercise:

Problem:

Fusion bombs use neutrons from their fission trigger to create tritium fuel in the reaction $n + ^6\text{Li} \rightarrow ^3\text{H} + ^4\text{He}$. What is the energy released by this reaction in MeV?

Solution:

4.781 MeV

Exercise:

Problem:

It is estimated that the total explosive yield of all the nuclear bombs in existence currently is about 4,000 MT.

(a) Convert this amount of energy to kilowatt-hours, noting that $1\ \text{kW} \cdot \text{h} = 3.60 \times 10^{6}\ \text{J}$.

(b) What would the monetary value of this energy be if it could be converted to electricity costing 10 cents per kW·h?

Exercise:

Problem:

A radiation-enhanced nuclear weapon (or neutron bomb) can have a smaller total yield and still produce more prompt radiation than a conventional nuclear bomb. This allows the use of neutron bombs to kill nearby advancing enemy forces with radiation without blowing up your own forces with the blast. For a 0.500-kT radiation-enhanced weapon and a 1.00-kT conventional nuclear bomb: (a) Compare the blast yields. (b) Compare the prompt radiation yields.

Solution:

(a) Blast yields $2.1 \times 10^{12}\ \text{J}$ to $8.4 \times 10^{11}\ \text{J}$, or 2.5 to 1, conventional to radiation enhanced.

(b) Prompt radiation yields $6.3 \times 10^{11}\ \text{J}$ to $2.1 \times 10^{11}\ \text{J}$, or 3 to 1, radiation enhanced to conventional.

Exercise:

Problem:

(a) How many ^{239}Pu nuclei must fission to produce a 20.0-kT yield, assuming 200 MeV per fission? (b) What is the mass of this much ^{239}Pu?

Exercise:

Problem:

Assume one-fourth of the yield of a typical 320-kT strategic bomb comes from fission reactions averaging 200 MeV and the remainder from fusion reactions averaging 20 MeV.

(a) Calculate the number of fissions and the approximate mass of uranium and plutonium fissioned, taking the average atomic mass to be 238.

(b) Find the number of fusions and calculate the approximate mass of fusion fuel, assuming an average total atomic mass of the two nuclei in each reaction to be 5.

(c) Considering the masses found, does it seem reasonable that some missiles could carry 10 warheads? Discuss, noting that the nuclear fuel is only a part of the mass of a warhead.

Solution:

(a) 1.1×10^{25} fissions , 4.4 kg

(b) 3.2×10^{26} fusions , 2.7 kg

(c) The nuclear fuel totals only 6 kg, so it is quite reasonable that some missiles carry 10 overheads. The mass of the fuel would only be 60 kg and therefore the mass of the 10 warheads, weighing about 10 times the nuclear fuel, would be only 1500 lbs. If the fuel for the missiles weighs 5 times the total weight of the warheads, the missile would weigh about 9000 lbs or 4.5 tons. This is not an unreasonable weight for a missile.

Exercise:

Problem:

This problem gives some idea of the magnitude of the energy yield of a small tactical bomb. Assume that half the energy of a 1.00-kT nuclear depth charge set off under an aircraft carrier goes into lifting it out of the water—that is, into gravitational potential energy. How high is the carrier lifted if its mass is 90,000 tons?

Exercise:

Problem:

It is estimated that weapons tests in the atmosphere have deposited approximately 9 MCi of ^{90}Sr on the surface of the earth. Find the mass of this amount of ^{90}Sr.

Solution:

$7 \times 10^4 \text{ g}$

Exercise:

Problem:

A 1.00-MT bomb exploded a few kilometers above the ground deposits 25.0% of its energy into radiant heat.

(a) Find the calories per cm^2 at a distance of 10.0 km by assuming a uniform distribution over a spherical surface of that radius.

(b) If this heat falls on a person’s body, what temperature increase does it cause in the affected tissue, assuming it is absorbed in a layer 1.00-cm deep?

Exercise:

Problem: Integrated Concepts

One scheme to put nuclear weapons to nonmilitary use is to explode them underground in a geologically stable region and extract the geothermal energy for electricity production. There was a total yield of about 4,000 MT in the combined arsenals in 2006. If 1.00 MT per day could be converted to electricity with an efficiency of 10.0%:

(a) What would the average electrical power output be?

(b) How many years would the arsenal last at this rate?

Solution:

(a) 4.86×10^9 W

(b) 11.0 y

Atomic Masses

Atomic Number, *Z*	**Name**	**Atomic Mass Number, *A***	**Symbol**	**Atomic Mass (u)**	**Percent Abundance or Decay Mode**	**Half-life, $t_{1/2}$**
0	neutron	1	n	1.008 665	β^-	10.37 min
1	Hydrogen	1	^1H	1.007 825	99.985%	
	Deuterium	2	^2H or D	2.014 102	0.015%	
	Tritium	3	^3H or T	3.016 050	β^-	12.33 y
2	Helium	3	^3He	3.016 030	$1.38 \times 10^{-4}\%$	
		4	^4He	4.002 603	$\approx 100\%$	
3	Lithium	6	^6Li	6.015 121	7.5%	
		7	^7Li	7.016 003	92.5%	
4	Beryllium	7	^7Be	7.016 928	EC	53.29 d
		9	^9Be	9.012 182	100%	
5	Boron	10	^{10}B	10.012 937	19.9%	
		11	^{11}B	11.009 305	80.1%	
6	Carbon	11	^{11}C	11.011 432	EC, β^+	
		12	^{12}C	12.000 000	98.90%	
		13	^{13}C	13.003 355	1.10%	

Atomic Number, Z	Name	Atomic Mass Number, A	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		14	${}^{14}\text{C}$	14.003 241	β^-	5730 y
7	Nitrogen	13	${}^{13}\text{N}$	13.005 738	β^+	9.96 min
		14	${}^{14}\text{N}$	14.003 074	99.63%	
		15	${}^{15}\text{N}$	15.000 108	0.37%	
8	Oxygen	15	${}^{15}\text{O}$	15.003 065	EC, β^+	122 s
		16	${}^{16}\text{O}$	15.994 915	99.76%	
		18	${}^{18}\text{O}$	17.999 160	0.200%	
9	Fluorine	18	${}^{18}\text{F}$	18.000 937	EC, β^+	1.83 h
		19	${}^{19}\text{F}$	18.998 403	100%	
10	Neon	20	${}^{20}\text{Ne}$	19.992 435	90.51%	
		22	${}^{22}\text{Ne}$	21.991 383	9.22%	
11	Sodium	22	${}^{22}\text{Na}$	21.994 434	β^+	2.602 y
		23	${}^{23}\text{Na}$	22.989 767	100%	
		24	${}^{24}\text{Na}$	23.990 961	β^-	14.96 h
12	Magnesium	24	${}^{24}\text{Mg}$	23.985 042	78.99%	
13	Aluminum	27	${}^{27}\text{Al}$	26.981 539	100%	
14	Silicon	28	${}^{28}\text{Si}$	27.976 927	92.23%	2.62h

Atomic Number, *Z*	**Name**	**Atomic Mass Number, *A***	**Symbol**	**Atomic Mass (u)**	**Percent Abundance or Decay Mode**	**Half-life, $t_{1/2}$**
		31	^{31}Si	30.975 362	β^-	
15	Phosphorus	31	^{31}P	30.973 762	100%	
		32	^{32}P	31.973 907	β^-	14.28 d
16	Sulfur	32	^{32}S	31.972 070	95.02%	
		35	^{35}S	34.969 031	β^-	87.4 d
17	Chlorine	35	^{35}Cl	34.968 852	75.77%	
		37	^{37}Cl	36.965 903	24.23%	
18	Argon	40	^{40}Ar	39.962 384	99.60%	
19	Potassium	39	^{39}K	38.963 707	93.26%	
		40	^{40}K	39.963 999	0.0117%, EC, β^-	$1.28 \times 10^9\text{y}$
20	Calcium	40	^{40}Ca	39.962 591	96.94%	
21	Scandium	45	^{45}Sc	44.955 910	100%	
22	Titanium	48	^{48}Ti	47.947 947	73.8%	
23	Vanadium	51	^{51}V	50.943 962	99.75%	
24	Chromium	52	^{52}Cr	51.940 509	83.79%	
25	Manganese	55	^{55}Mn	54.938 047	100%	
26	Iron	56	^{56}Fe	55.934 939	91.72%	

Atomic Number, *Z*	**Name**	**Atomic Mass Number, *A***	**Symbol**	**Atomic Mass (u)**	**Percent Abundance or Decay Mode**	**Half-life, $t_{1/2}$**
27	Cobalt	59	^{59}Co	58.933 198	100%	
		60	^{60}Co	59.933 819	β^-	5.271 y
28	Nickel	58	^{58}Ni	57.935 346	68.27%	
		60	^{60}Ni	59.930 788	26.10%	
29	Copper	63	^{63}Cu	62.939 598	69.17%	
		65	^{65}Cu	64.927 793	30.83%	
30	Zinc	64	^{64}Zn	63.929 145	48.6%	
		66	^{66}Zn	65.926 034	27.9%	
31	Gallium	69	^{69}Ga	68.925 580	60.1%	
32	Germanium	72	^{72}Ge	71.922 079	27.4%	
		74	^{74}Ge	73.921 177	36.5%	
33	Arsenic	75	^{75}As	74.921 594	100%	
34	Selenium	80	^{80}Se	79.916 520	49.7%	
35	Bromine	79	^{79}Br	78.918 336	50.69%	
36	Krypton	84	^{84}Kr	83.911 507	57.0%	
37	Rubidium	85	^{85}Rb	84.911 794	72.17%	
38	Strontium	86	^{86}Sr	85.909 267	9.86%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		88	^{88}Sr	87.905 619	82.58%	
		90	^{90}Sr	89.907 738	β^-	28.8 y
39	Yttrium	89	^{89}Y	88.905 849	100%	
		90	^{90}Y	89.907 152	β^-	64.1 h
40	Zirconium	90	^{90}Zr	89.904 703	51.45%	
41	Niobium	93	^{93}Nb	92.906 377	100%	
42	Molybdenum	98	^{98}Mo	97.905 406	24.13%	
43	Technetium	98	^{98}Tc	97.907 215	β^-	$4.2 \times 10^6\text{y}$
44	Ruthenium	102	^{102}Ru	101.904 348	31.6%	
45	Rhodium	103	^{103}Rh	102.905 500	100%	
46	Palladium	106	^{106}Pd	105.903 478	27.33%	
47	Silver	107	^{107}Ag	106.905 092	51.84%	
		109	^{109}Ag	108.904 757	48.16%	
48	Cadmium	114	^{114}Cd	113.903 357	28.73%	
49	Indium	115	^{115}In	114.903 880	95.7%, β^-	$4.4 \times 10^{14}\text{y}$
50	Tin	120	^{120}Sn	119.902 200	32.59%	
51	Antimony	121	^{121}Sb	120.903 821	57.3%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
52	Tellurium	130	^{130}Te	129.906 229	33.8%, β^-	$2.5 \times 10^{21}\text{y}$
53	Iodine	127	^{127}I	126.904 473	100%	
		131	^{131}I	130.906 114	β^-	8.040 d
54	Xenon	132	^{132}Xe	131.904 144	26.9%	
		136	^{136}Xe	135.907 214	8.9%	
55	Cesium	133	^{133}Cs	132.905 429	100%	
		134	^{134}Cs	133.906 696	EC, β^-	2.06 y
56	Barium	137	^{137}Ba	136.905 812	11.23%	
		138	^{138}Ba	137.905 232	71.70%	
57	Lanthanum	139	^{139}La	138.906 346	99.91%	
58	Cerium	140	^{140}Ce	139.905 433	88.48%	
59	Praseodymium	141	^{141}Pr	140.907 647	100%	
60	Neodymium	142	^{142}Nd	141.907 719	27.13%	
61	Promethium	145	^{145}Pm	144.912 743	EC, α	17.7 y
62	Samarium	152	^{152}Sm	151.919 729	26.7%	
63	Europium	153	^{153}Eu	152.921 225	52.2%	
64	Gadolinium	158	^{158}Gd	157.924 099	24.84%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
65	Terbium	159	^{159}Tb	158.925 342	100%	
66	Dysprosium	164	^{164}Dy	163.929 171	28.2%	
67	Holmium	165	^{165}Ho	164.930 319	100%	
68	Erbium	166	^{166}Er	165.930 290	33.6%	
69	Thulium	169	^{169}Tm	168.934 212	100%	
70	Ytterbium	174	^{174}Yb	173.938 859	31.8%	
71	Lutecium	175	^{175}Lu	174.940 770	97.41%	
72	Hafnium	180	^{180}Hf	179.946 545	35.10%	
73	Tantalum	181	^{181}Ta	180.947 992	99.98%	
74	Tungsten	184	^{184}W	183.950 928	30.67%	
75	Rhenium	187	^{187}Re	186.955 744	62.6%, β^-	$4.6 \times 10^{10}\text{y}$
76	Osmium	191	^{191}Os	190.960 920	β^-	15.4 d
		192	^{192}Os	191.961 467	41.0%	
77	Iridium	191	^{191}Ir	190.960 584	37.3%	
		193	^{193}Ir	192.962 917	62.7%	
78	Platinum	195	^{195}Pt	194.964 766	33.8%	
79	Gold	197	^{197}Au	196.966 543	100%	

Atomic Number, *Z*	Name	Atomic Mass Number, *A*	Symbol	Atomic Mass (u)	Percent Abundance or Decay Mode	Half-life, $t_{1/2}$
		198	$^{198}\mathrm{Au}$	197.968 217	β^-	2.696 d
80	Mercury	199	$^{199}\mathrm{Hg}$	198.968 253	16.87%	
		202	$^{202}\mathrm{Hg}$	201.970 617	29.86%	
81	Thallium	205	$^{205}\mathrm{Tl}$	204.974 401	70.48%	
82	Lead	206	$^{206}\mathrm{Pb}$	205.974 440	24.1%	
		207	$^{207}\mathrm{Pb}$	206.975 872	22.1%	
		208	$^{208}\mathrm{Pb}$	207.976 627	52.4%	
		210	$^{210}\mathrm{Pb}$	209.984 163	α, β^-	22.3 y
		211	$^{211}\mathrm{Pb}$	210.988 735	β^-	36.1 min
		212	$^{212}\mathrm{Pb}$	211.991 871	β^-	10.64 h
83	Bismuth	209	$^{209}\mathrm{Bi}$	208.980 374	100%	
		211	$^{211}\mathrm{Bi}$	210.987 255	α, β^-	2.14 min
84	Polonium	210	$^{210}\mathrm{Po}$	209.982 848	α	138.38 d
85	Astatine	218	$^{218}\mathrm{At}$	218.008 684	α, β^-	1.6 s
86	Radon	222	$^{222}\mathrm{Rn}$	222.017 570	α	3.82 d
87	Francium	223	$^{223}\mathrm{Fr}$	223.019 733	α, β^-	21.8 min
88	Radium	226	$^{226}\mathrm{Ra}$	226.025 402	α	1.60×10^3y

Atomic Number, *Z*	**Name**	**Atomic Mass Number, *A***	**Symbol**	**Atomic Mass (u)**	**Percent Abundance or Decay Mode**	**Half-life, $t_{1/2}$**
89	Actinium	227	^{227}Ac	227.027 750	α, β^-	21.8 y
90	Thorium	228	^{228}Th	228.028 715	α	1.91 y
		232	^{232}Th	232.038 054	100%, α	$1.41 \times 10^{10}\text{y}$
91	Protactinium	231	^{231}Pa	231.035 880	α	$3.28 \times 10^{4}\text{y}$
92	Uranium	233	^{233}U	233.039 628	α	$1.59 \times 10^{3}\text{y}$
		235	^{235}U	235.043 924	0.720%, α	$7.04 \times 10^{8}\text{y}$
		236	^{236}U	236.045 562	α	$2.34 \times 10^{7}\text{y}$
		238	^{238}U	238.050 784	99.2745%, α	$4.47 \times 10^{9}\text{y}$
		239	^{239}U	239.054 289	β^-	23.5 min
93	Neptunium	239	^{239}Np	239.052 933	β^-	2.355 d
94	Plutonium	239	^{239}Pu	239.052 157	α	$2.41 \times 10^{4}\text{y}$
95	Americium	243	^{243}Am	243.061 375	α, fission	$7.37 \times 10^{3}\text{y}$
96	Curium	245	^{245}Cm	245.065 483	α	$8.50 \times 10^{3}\text{y}$
97	Berkelium	247	^{247}Bk	247.070 300	α	$1.38 \times 10^{3}\text{y}$
98	Californium	249	^{249}Cf	249.074 844	α	351 y
99	Einsteinium	254	^{254}Es	254.088 019	α, β^-	276 d
100	Fermium	253	^{253}Fm	253.085 173	EC, α	3.00 d

Atomic Number, *Z*	**Name**	**Atomic Mass Number, *A***	**Symbol**	**Atomic Mass (u)**	**Percent Abundance or Decay Mode**	**Half-life, $t_{1/2}$**
101	Mendelevium	255	^{255}Md	255.091 081	EC, α	27 min
102	Nobelium	255	^{255}No	255.093 260	EC, α	3.1 min
103	Lawrencium	257	^{257}Lr	257.099 480	EC, α	0.646 s
104	Rutherfordium	261	^{261}Rf	261.108 690	α	1.08 min
105	Dubnium	262	^{262}Db	262.113 760	α, fission	34 s
106	Seaborgium	263	^{263}Sg	263.11 86	α, fission	0.8 s
107	Bohrium	262	^{262}Bh	262.123 1	α	0.102 s
108	Hassium	264	^{264}Hs	264.128 5	α	0.08 ms
109	Meitnerium	266	^{266}Mt	266.137 8	α	3.4 ms

Atomic Masses

Selected Radioactive Isotopes

Decay modes are α, β^-, β^+, electron capture (EC) and isomeric transition (IT). EC results in the same daughter nucleus as would β^+ decay. IT is a transition from a metastable excited state. Energies for $\beta^\pm$ decays are the maxima; average energies are roughly one-half the maxima.

Isotope	***t* 1/2**	**DecayMode(s)**	**Energy(MeV)**	**Percent**		**γ-Ray Energy(MeV)**
^{3}H	12.33 y	β^-	0.0186	100%		
^{14}C	5730 y	β^-	0.156	100%		
^{13}N	9.96 min	β^+	1.20	100%		
^{22}Na	2.602 y	β^+	0.55	90%	γ	1.27
^{32}P	14.28 d	β^-	1.71	100%		
^{35}S	87.4 d	β^-	0.167	100%		
^{36}Cl	3.00×10^5y	β^-	0.710	100%		
^{40}K	1.28×10^9y	β^-	1.31	89%		
^{43}K	22.3 h	β^-	0.827	87%	γs	0.373
						0.618
^{45}Ca	165 d	β^-	0.257	100%		
^{51}Cr	27.70 d	EC			γ	0.320
^{52}Mn	5.59d	β^+	3.69	28%	γs	1.33
						1.43
^{52}Fe	8.27 h	β^+	1.80	43%		0.169
						0.378
^{59}Fe	44.6 d	β^-s	0.273	45%	γs	1.10
			0.466	55%		1.29
^{60}Co	5.271 y	β^-	0.318	100%	γs	1.17
						1.33
^{65}Zn	244.1 d	EC			γ	1.12

Isotope	$t_{1/2}$	DecayMode(s)	Energy(MeV)	Percent		γ-Ray Energy(MeV)
$^{67}\mathrm{Ga}$	78.3 h	EC			γs	0.0933
						0.185
						0.300
						others
$^{75}\mathrm{Se}$	118.5 d	EC			γs	0.121
						0.136
						0.265
						0.280
						others
$^{86}\mathrm{Rb}$	18.8 d	β^-s	0.69	9%	γ	1.08
			1.77	91%		
$^{85}\mathrm{Sr}$	64.8 d	EC			γ	0.514
$^{90}\mathrm{Sr}$	28.8 y	β^-	0.546	100%		
$^{90}\mathrm{Y}$	64.1 h	β^-	2.28	100%		
$^{99\mathrm{m}}\mathrm{Tc}$	6.02 h	IT			γ	0.142
$^{113\mathrm{m}}\mathrm{In}$	99.5 min	IT			γ	0.392
$^{123}\mathrm{I}$	13.0 h	EC			γ	0.159
$^{131}\mathrm{I}$	8.040 d	β^-s	0.248	7%	γs	0.364
			0.607	93%		others
			others			
$^{129}\mathrm{Cs}$	32.3 h	EC			γs	0.0400
						0.372
						0.411
						others
$^{137}\mathrm{Cs}$	30.17 y	β^-s	0.511	95%	γ	0.662
			1.17	5%		

Isotope	***t* 1/2**	**DecayMode(s)**	**Energy(MeV)**	**Percent**		**γ-Ray Energy(MeV)**
^{140}Ba	12.79 d	β^-	1.035	$\approx 100\%$	γs	0.030
						0.044
						0.537
						others
^{198}Au	2.696 d	β^-	1.161	$\approx 100\%$	γ	0.412
^{197}Hg	64.1 h	EC			γ	0.0733
^{210}Po	138.38 d	α	5.41	100%		
^{226}Ra	1.60×10^3y	αs	4.68	5%	γ	0.186
			4.87	95%		
^{235}U	7.038×10^8y	α	4.68	$\approx 100\%$	γs	numerous
^{238}U	4.468×10^9y	αs	4.22	23%	γ	0.050
			4.27	77%		
^{237}Np	2.14×10^6y	αs	numerous		γs	numerous
			4.96 (max.)			
^{239}Pu	2.41×10^4y	αs	5.19	11%	γs	7.5×10^{-5}
			5.23	15%		0.013
			5.24	73%		0.052
						others
^{243}Am	7.37×10^3y	αs	Max. 5.44		γs	0.075
			5.37	88%		others
			5.32	11%		
			others			

Selected Radioactive Isotopes

Useful Information

This appendix is broken into several tables.

- [link], Important Constants
- [link], Submicroscopic Masses
- [link], Solar System Data
- [link], Metric Prefixes for Powers of Ten and Their Symbols
- [link], The Greek Alphabet
- [link], SI units
- [link], Selected British Units
- [link], Other Units
- [link], Useful Formulae

Symbol	Meaning	Best Value	Approximate Value
c	Speed of light in vacuum	$2.99792458 \times 10^{8}\,\mathrm{m/s}$	$3.00 \times 10^{8}\,\mathrm{m/s}$
G	Gravitational constant	$6.67408(31) \times 10^{-11}\,\mathrm{N\cdot m^2/kg^2}$	$6.67 \times 10^{-11}\,\mathrm{N\cdot m^2/kg^2}$
N_A	Avogadro's number	$6.02214129(27) \times 10^{23}$	6.02×10^{23}
k	Boltzmann's constant	$1.3806488(13) \times 10^{-23}\,\mathrm{J/K}$	$1.38 \times 10^{-23}\,\mathrm{J/K}$
R	Gas constant	$8.3144621(75)\,\mathrm{J/mol\cdot K}$	$8.31\,\mathrm{J/mol\cdot K} = 1.99\,\mathrm{cal/mol\cdot K} =$
σ	Stefan-Boltzmann constant	$5.670373(21) \times 10^{-8}\,\mathrm{W/m^2\cdot K}$	$5.67 \times 10^{-8}\,\mathrm{W/m^2\cdot K}$
k	Coulomb force constant	$8.987551788... \times 10^{9}\,\mathrm{N\cdot m^2/C^2}$	$8.99 \times 10^{9}\,\mathrm{N\cdot m^2/C^2}$
q_e	Charge on electron	$-1.602176565(35) \times 10^{-19}\,\mathrm{C}$	$-1.60 \times 10^{-19}\,\mathrm{C}$
ε_0	Permittivity of free space	$8.854187817... \times 10^{-12}\,\mathrm{C^2/N\cdot m^2}$	$8.85 \times 10^{-12}\,\mathrm{C^2/N\cdot m^2}$
μ_0	Permeability of free space	$4\pi \times 10^{-7}\,\mathrm{T\cdot m/A}$	$1.26 \times 10^{-6}\,\mathrm{T\cdot m/A}$
h	Planck's constant	$6.62606957(29) \times 10^{-34}\,\mathrm{J\cdot s}$	$6.63 \times 10^{-34}\,\mathrm{J\cdot s}$

Important Constants[footnote]

Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, an www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Nu are exact as defined.

Symbol	Meaning	Best Value	Approximate Value
m_e	Electron mass	$9.10938291(40) \times 10^{-31}\text{kg}$	$9.11 \times 10^{-31}\text{kg}$
m_p	Proton mass	$1.672621777(74) \times 10^{-27}\text{kg}$	$1.6726 \times 10^{-27}\text{kg}$
m_n	Neutron mass	$1.674927351(74) \times 10^{-27}\text{kg}$	$1.6749 \times 10^{-27}\text{kg}$
u	Atomic mass unit	$1.660538921(73) \times 10^{-27}\text{kg}$	$1.6605 \times 10^{-27}\text{kg}$

Submicroscopic Masses[footnote]
Stated values are according to the National Institute of Standards and Technology Reference on Constants, Units, and Uncertainty, www.physics.nist.gov/cuu (accessed May 18, 2012). Values in parentheses are the uncertainties in the last digits. Numbers without uncertainties are exact as defined.

Sun	mass	$1.99 \times 10^{30}\text{kg}$
	average radius	$6.96 \times 10^{8}\text{m}$
	Earth-sun distance (average)	$1.496 \times 10^{11}\text{m}$
Earth	mass	$5.9736 \times 10^{24}\text{kg}$
	average radius	$6.376 \times 10^{6}\text{m}$
	orbital period	$3.16 \times 10^{7}\text{s}$

Moon	mass	$7.35 \times 10^{22} \mathrm{kg}$
	average radius	$1.74 \times 10^{6} \mathrm{m}$
	orbital period (average)	$2.36 \times 10^{6} \mathrm{s}$
	Earth-moon distance (average)	$3.84 \times 10^{8} \mathrm{m}$

Solar System Data

Prefix	**Symbol**	**Value**	**Prefix**	**Symbol**	**Value**
tera	T	10^{12}	deci	d	10^{-1}
giga	G	10^{9}	centi	c	10^{-2}
mega	M	10^{6}	milli	m	10^{-3}
kilo	k	10^{3}	micro	μ	10^{-6}
hecto	h	10^{2}	nano	n	10^{-9}
deka	da	10^{1}	pico	p	10^{-12}
—	—	$10^{0} (= 1)$	femto	f	10^{-15}

Metric Prefixes for Powers of Ten and Their Symbols

Alpha	A	α	Eta	H	η	Nu	N	ν	Tau	T	τ
Beta	B	β	Theta	Θ	θ	Xi	Ξ	ξ	Upsilon	Υ	υ
Gamma	Γ	γ	Iota	I	ι	Omicron	O	o	Phi	Φ	ϕ
Delta	Δ	δ	Kappa	K	κ	Pi	Π	π	Chi	X	χ

Epsilon	E	ε	Lambda	Λ	λ	Rho	P	ρ	Psi	Ψ	ψ
Zeta	Z	ζ	Mu	M	μ	Sigma	Σ	σ	Omega	Ω	ω

The Greek Alphabet

	Entity	**Abbreviation**	**Name**
Fundamental units	Length	m	meter
	Mass	kg	kilogram
	Time	s	second
	Current	A	ampere
Supplementary unit	Angle	rad	radian
Derived units	Force	$\mathrm{N} = \mathrm{kg} \cdot \mathrm{m}/\mathrm{s}^2$	newton
	Energy	$\mathrm{J} = \mathrm{kg} \cdot \mathrm{m}^2/\mathrm{s}^2$	joule
	Power	$\mathrm{W} = \mathrm{J}/\mathrm{s}$	watt
	Pressure	$\mathrm{Pa} = \mathrm{N}/\mathrm{m}^2$	pascal
	Frequency	$\mathrm{Hz} = 1/\mathrm{s}$	hertz
	Electronic potential	$\mathrm{V} = \mathrm{J}/\mathrm{C}$	volt
	Capacitance	$\mathrm{F} = \mathrm{C}/\mathrm{V}$	farad
	Charge	$\mathrm{C} = \mathrm{s} \cdot \mathrm{A}$	coulomb
	Resistance	$\Omega = \mathrm{V}/\mathrm{A}$	ohm

	Entity	**Abbreviation**	**Name**
	Magnetic field	$\mathrm{T} = \mathrm{N}/(\mathrm{A} \cdot \mathrm{m})$	tesla
	Nuclear decay rate	$\mathrm{Bq} = 1/\mathrm{s}$	becquerel

SI Units

Length	$1\ \text{inch (in.)} = 2.54\ \text{cm (exactly)}$
	$1\ \text{foot (ft)} = 0.3048\ \text{m}$
	$1\ \text{mile (mi)} = 1.609\ \text{km}$
Force	$1\ \text{pound (lb)} = 4.448\ \text{N}$
Energy	$1\ \text{British thermal unit (Btu)} = 1.055 \times 10^3\ \text{J}$
Power	$1\ \text{horsepower (hp)} = 746\ \text{W}$
Pressure	$1\ \text{lb/in}^2 = 6.895 \times 10^3\ \text{Pa}$

Selected British Units

Length	$1\ \text{light year (ly)} = 9.46 \times 10^{15}\ \text{m}$
	$1\ \text{astronomical unit (au)} = 1.50 \times 10^{11}\ \text{m}$
	$1\ \text{nautical mile} = 1.852\ \text{km}$
	$1\ \text{angstrom(Å)}\ = 10^{-10}\ \text{m}$
Area	$1\ \text{acre (ac)} = 4.05 \times 10^3\ \text{m}^2$
	$1\ \text{square foot (ft}^2) = 9.29 \times 10^{-2}\ \text{m}^2$
	$1\ \text{barn}\ (b) = 10^{-28}\ \text{m}^2$
Volume	$1\ \text{liter}\ (L) = 10^{-3}\ \text{m}^3$

	$1\,\text{U.S. gallon (gal)} = 3.785 \times 10^{-3}\,\text{m}^3$
Mass	$1\,\text{solar mass} = 1.99 \times 10^{30}\,\text{kg}$
	$1\,\text{metric ton} = 10^3\,\text{kg}$
	$1\,\text{atomic mass unit}\,(u) = 1.6605 \times 10^{-27}\,\text{kg}$
Time	$1\,\text{year}\,(y) = 3.16 \times 10^7\,\text{s}$
	$1\,\text{day}\,(d) = 86\ 400\,\text{s}$
Speed	$1\,\text{mile per hour (mph)} = 1.609\,\text{km/h}$
	$1\,\text{nautical mile per hour (naut)} = 1.852\,\text{km/h}$
Angle	$1\,\text{degree}\,(^\circ) = 1.745 \times 10^{-2}\,\text{rad}$
	$1\,\text{minute of arc}\,(') = 1/60\,\text{degree}$
	$1\,\text{second of arc}\,('') = 1/60\,\text{minute of arc}$
	$1\,\text{grad} = 1.571 \times 10^{-2}\,\text{rad}$
Energy	$1\,\text{kiloton TNT (kT)} = 4.2 \times 10^{12}\,\text{J}$
	$1\,\text{kilowatt hour}\,(\text{kW} \cdot h) = 3.60 \times 10^6\,\text{J}$
	$1\,\text{food calorie (kcal)} = 4186\,\text{J}$
	$1\,\text{calorie (cal)} = 4.186\,\text{J}$
	$1\,\text{electron volt (eV)} = 1.60 \times 10^{-19}\,\text{J}$
Pressure	$1\,\text{atmosphere (atm)} = 1.013 \times 10^5\,\text{Pa}$
	$1\,\text{millimeter of mercury (mm Hg)} = 133.3\,\text{Pa}$
	$1\,\text{torricelli (torr)} = 1\,\text{mm Hg} = 133.3\,\text{Pa}$
Nuclear decay rate	$1\,\text{curie (Ci)} = 3.70 \times 10^{10}\,\text{Bq}$

Other Units

Circumference of a circle with radius r or diameter d	$C = 2\pi r = \pi d$
Area of a circle with radius r or diameter d	$A = \pi r^2 = \pi d^2/4$
Area of a sphere with radius r	$A = 4\pi r^2$

Volume of a sphere with radius r	$V = (4/3)\ \pi r^3$

Useful Formulae

Glossary of Key Symbols and Notation

In this glossary, key symbols and notation are briefly defined.

Symbol	Definition
any symbol	average (indicated by a bar over a symbol—e.g., v is average velocity)
$^{\circ}\text{C}$	Celsius degree
$^{\circ}\text{F}$	Fahrenheit degree
$//$	parallel
$\perp$	perpendicular
$\propto$	proportional to
$\pm$	plus or minus

Symbol	Definition
$_0$	zero as a subscript denotes an initial value
α	alpha rays
α	angular acceleration
α	temperature coefficient(s) of resistivity
β	beta rays
β	sound level
β	volume coefficient of expansion
β^-	electron emitted in nuclear beta decay
β^+	positron decay
γ	gamma rays

Symbol	Definition
γ	surface tension
$\gamma = 1/\sqrt{1 - v^2/c^2}$	a constant used in relativity
Δ	change in whatever quantity follows
δ	uncertainty in whatever quantity follows
$\mathit{\Delta} E$	change in energy between the initial and final orbits of an electron in an atom
$\mathit{\Delta} E$	uncertainty in energy
$\mathit{\Delta} m$	difference in mass between initial and final products
$\mathit{\Delta} N$	number of decays that occur
$\mathit{\Delta} p$	change in momentum

Symbol	Definition
Δp	uncertainty in momentum
$\Delta \mathrm{PE}_{\mathrm{g}}$	change in gravitational potential energy
$\Delta \theta$	rotation angle
Δs	distance traveled along a circular path
Δt	uncertainty in time
Δt_0	proper time as measured by an observer at rest relative to the process
ΔV	potential difference
Δx	uncertainty in position
ε_0	permittivity of free space
η	viscosity

Symbol	Definition
θ	angle between the force vector and the displacement vector
θ	angle between two lines
θ	contact angle
θ	direction of the resultant
θ_b	Brewster's angle
θ_c	critical angle
κ	dielectric constant
λ	decay constant of a nuclide
λ	wavelength
λ_n	wavelength in a medium

Symbol	Definition
μ_0	permeability of free space
μ_k	coefficient of kinetic friction
μ_s	coefficient of static friction
v_e	electron neutrino
π^+	positive pion
π^-	negative pion
π^0	neutral pion
ρ	density
ρ_c	critical density, the density needed to just halt universal expansion
ρ_fl	fluid density

Symbol	**Definition**
ρ_{obj}	average density of an object
ρ/ρ_{w}	specific gravity
τ	characteristic time constant for a resistance and inductance (RL) or resistance and capacitance (RC) circuit
τ	characteristic time for a resistor and capacitor (RC) circuit
τ	torque
Υ	upsilon meson
Φ	magnetic flux
ϕ	phase angle
Ω	ohm (unit)
ω	angular velocity

Symbol	Definition
A	ampere (current unit)
A	area
A	cross-sectional area
A	total number of nucleons
a	acceleration
a_{B}	Bohr radius
a_{c}	centripetal acceleration
a_{t}	tangential acceleration
AC	alternating current
AM	amplitude modulation

Symbol	**Definition**
atm	atmosphere
B	baryon number
B	blue quark color
B	antiblue (yellow) antiquark color
b	quark flavor bottom or beauty
B	bulk modulus
B	magnetic field strength
$\mathrm{B}_{\mathrm{int}}$	electron’s intrinsic magnetic field
$\mathrm{B}_{\mathrm{orb}}$	orbital magnetic field
BE	binding energy of a nucleus—it is the energy required to completely disassemble it into separate protons and neutrons

Symbol	Definition
BE/A	binding energy per nucleon
Bq	becquerel—one decay per second
C	capacitance (amount of charge stored per volt)
C	coulomb (a fundamental SI unit of charge)
C_{p}	total capacitance in parallel
C_{s}	total capacitance in series
CG	center of gravity
CM	center of mass
c	quark flavor charm
c	specific heat

Symbol	**Definition**
c	speed of light
Cal	kilocalorie
cal	calorie
COP_{hp}	heat pump’s coefficient of performance
COP_{ref}	coefficient of performance for refrigerators and air conditioners
$\cos\theta$	cosine
$\cot\theta$	cotangent
$\csc\theta$	cosecant
D	diffusion constant
d	displacement

Symbol	Definition
d	quark flavor down
dB	decibel
d_{i}	distance of an image from the center of a lens
d_{o}	distance of an object from the center of a lens
DC	direct current
E	electric field strength
ε	emf (voltage) or Hall electromotive force
emf	electromotive force
E	energy of a single photon
E	nuclear reaction energy

Symbol	**Definition**
E	relativistic total energy
E	total energy
E_0	ground state energy for hydrogen
E_0	rest energy
EC	electron capture
E_{cap}	energy stored in a capacitor
Eff	efficiency—the useful work output divided by the energy input
Eff_{C}	Carnot efficiency
E_{in}	energy consumed (food digested in humans)
E_{ind}	energy stored in an inductor

Symbol	Definition
E_{out}	energy output
e	emissivity of an object
e^+	antielectron or positron
eV	electron volt
F	farad (unit of capacitance, a coulomb per volt)
F	focal point of a lens
	force
F	magnitude of a force
F	restoring force
F_{B}	buoyant force

Symbol	Definition
F_{c}	centripetal force
F_{i}	force input
$_{\mathrm{net}}$	net force
F_{o}	force output
FM	frequency modulation
f	focal length
f	frequency
f_0	resonant frequency of a resistance, inductance, and capacitance (RLC) series circuit
f_0	threshold frequency for a particular material (photoelectric effect)

Symbol	Definition
f_1	fundamental
f_2	first overtone
f_3	second overtone
f_B	beat frequency
f_k	magnitude of kinetic friction
f_s	magnitude of static friction
G	gravitational constant
G	green quark color
G	antigreen (magenta) antiquark color

Symbol	**Definition**
g	acceleration due to gravity
g	gluons (carrier particles for strong nuclear force)
h	change in vertical position
h	height above some reference point
h	maximum height of a projectile
h	Planck's constant
hf	photon energy
h_{i}	height of the image
h_{o}	height of the object
I	electric current

Symbol	Definition
I	intensity
I	intensity of a transmitted wave
I	moment of inertia (also called rotational inertia)
I_0	intensity of a polarized wave before passing through a filter
I_{ave}	average intensity for a continuous sinusoidal electromagnetic wave
I_{rms}	average current
J	joule
J/Ψ	Joules/psi meson
K	kelvin
k	Boltzmann constant

Symbol	Definition
k	force constant of a spring
K_α	x rays created when an electron falls into an $n=1$ shell vacancy from the $n=3$ shell
K_β	x rays created when an electron falls into an $n=2$ shell vacancy from the $n=3$ shell
kcal	kilocalorie
KE	translational kinetic energy
$\mathrm{KE}+\mathrm{PE}$	mechanical energy
KE_e	kinetic energy of an ejected electron
$\mathrm{KE}_{\mathrm{rel}}$	relativistic kinetic energy
$\mathrm{KE}_{\mathrm{rot}}$	rotational kinetic energy
KE	thermal energy

Symbol	Definition
kg	kilogram (a fundamental SI unit of mass)
L	angular momentum
L	liter
L	magnitude of angular momentum
L	self-inductance
ℓ	angular momentum quantum number
L_α	x rays created when an electron falls into an $n=2$ shell from the $n=3$ shell
L_e	electron total family number
L_μ	muon family total number
L_τ	tau family total number

Symbol	Definition
L_f	heat of fusion
L_f and L_v	latent heat coefficients
L_orb	orbital angular momentum
L_s	heat of sublimation
L_v	heat of vaporization
L_z	*z* - component of the angular momentum
M	angular magnification
M	mutual inductance
m	indicates metastable state
m	magnification

Symbol	Definition
m	mass
m	mass of an object as measured by a person at rest relative to the object
m	meter (a fundamental SI unit of length)
m	order of interference
m	overall magnification (product of the individual magnifications)
$m\left({}^{A}\mathrm{X}\right)$	atomic mass of a nuclide
MA	mechanical advantage
m_{e}	magnification of the eyepiece
m_e	mass of the electron
m_ℓ	angular momentum projection quantum number

Symbol	Definition
m_n	mass of a neutron
m_{o}	magnification of the objective lens
mol	mole
m_p	mass of a proton
m_{s}	spin projection quantum number
N	magnitude of the normal force
N	newton
	normal force
N	number of neutrons
n	index of refraction

Symbol	Definition
n	number of free charges per unit volume
N_{A}	Avogadro's number
N_{r}	Reynolds number
$\mathrm{N} \cdot \mathrm{m}$	newton-meter (work-energy unit)
$\mathrm{N} \cdot \mathrm{m}$	newtons times meters (SI unit of torque)
OE	other energy
P	power
P	power of a lens
P	pressure
	momentum

Symbol	Definition
p	momentum magnitude
p	relativistic momentum
$_{\text{tot}}$	total momentum
$_{\text{tot}}$	total momentum some time later
P_{abs}	absolute pressure
P_{atm}	atmospheric pressure
P_{atm}	standard atmospheric pressure
PE	potential energy
PE_{el}	elastic potential energy
PE_{elec}	electric potential energy

Symbol	Definition
PE_s	potential energy of a spring
P_g	gauge pressure
P_in	power consumption or input
P_out	useful power output going into useful work or a desired, form of energy
Q	latent heat
Q	net heat transferred into a system
Q	flow rate—volume per unit time flowing past a point
$+Q$	positive charge
$-Q$	negative charge

Symbol	Definition
q	electron charge
q_p	charge of a proton
q	test charge
QF	quality factor
R	activity, the rate of decay
R	radius of curvature of a spherical mirror
R	red quark color
R	antired (cyan) quark color
R	resistance
R	resultant or total displacement

Symbol	Definition
R	Rydberg constant
R	universal gas constant
r	distance from pivot point to the point where a force is applied
r	internal resistance
$r_\perp$	perpendicular lever arm
r	radius of a nucleus
r	radius of curvature
r	resistivity
r or rad	radiation dose unit
rem	roentgen equivalent man

Symbol	Definition
rad	radian
RBE	relative biological effectiveness
RC	resistor and capacitor circuit
rms	root mean square
r_n	radius of the *n*th H-atom orbit
R_{p}	total resistance of a parallel connection
R_{s}	total resistance of a series connection
R_{s}	Schwarzschild radius
S	entropy
	intrinsic spin (intrinsic angular momentum)

Symbol	Definition
S	magnitude of the intrinsic (internal) spin angular momentum
S	shear modulus
S	strangeness quantum number
s	quark flavor strange
s	second (fundamental SI unit of time)
s	spin quantum number
	total displacement
$\sec\theta$	secant
$\sin\theta$	sine
s_z	*z*-component of spin angular momentum

Symbol	Definition
T	period—time to complete one oscillation
T	temperature
T_{c}	critical temperature—temperature below which a material becomes a superconductor
T	tension
T	tesla (magnetic field strength *B*)
t	quark flavor top or truth
t	time
$t_{1/2}$	half-life—the time in which half of the original nuclei decay
$\tan\theta$	tangent
U	internal energy

Symbol	Definition
u	quark flavor up
u	unified atomic mass unit
	velocity of an object relative to an observer
	velocity relative to another observer
V	electric potential
V	terminal voltage
V	volt (unit)
V	volume
	relative velocity between two observers
v	speed of light in a material

Symbol	Definition
	velocity
	average fluid velocity
$V_{\text{B}} - V_{\text{A}}$	change in potential
$_{\text{d}}$	drift velocity
V_{p}	transformer input voltage
V_{rms}	rms voltage
V_{s}	transformer output voltage
$_{\text{tot}}$	total velocity
v_{w}	propagation speed of sound or other wave
$_{\text{w}}$	wave velocity

Symbol	Definition
W	work
W	net work done by a system
W	watt
w	weight
w_{fl}	weight of the fluid displaced by an object
W_{c}	total work done by all conservative forces
W_{nc}	total work done by all nonconservative forces
W_{out}	useful work output
X	amplitude
X	symbol for an element

Symbol	Definition
${}_{A}^{Z}X_N$	notation for a particular nuclide
x	deformation or displacement from equilibrium
x	displacement of a spring from its undeformed position
x	horizontal axis
X_{C}	capacitive reactance
X_{L}	inductive reactance
x_{rms}	root mean square diffusion distance
y	vertical axis
Y	elastic modulus or Young's modulus
Z	atomic number (number of protons in a nucleus)

Symbol	Definition
Z	impedance